

# Transformer Based Multi-Source Domain Adaptation

Dustin Wright and Isabelle Augenstein

Dept. of Computer Science

University of Copenhagen

Denmark

{dw|augenstein}@di.ku.dk

## Abstract

In practical machine learning settings, the data on which a model must make predictions often come from a different distribution than the data it was trained on. Here, we investigate the problem of *unsupervised multi-source domain adaptation*, where a model is trained on labelled data from multiple source domains and must make predictions on a domain for which no labelled predictions has been seen. Prior work with CNNs and RNNs has demonstrated the benefit of mixture of experts, where the predictions of multiple domain expert classifiers are combined; as well as domain adversarial training, to induce a domain agnostic representation space. Inspired by this, we investigate how such methods can be effectively applied to large pretrained transformer models. We find that domain adversarial training has an effect on the learned representations of these models while having little effect on their performance, suggesting that large transformer-based models are already relatively robust across domains. Additionally, we show that mixture of experts leads to significant performance improvements by comparing several variants of mixing functions, including one novel mixture based on attention. Finally, we demonstrate that the predictions of large pretrained transformer based domain experts are highly homogenous, making it challenging to learn effective functions for mixing their predictions.

## 1 Introduction

Machine learning practitioners are often faced with the problem of evolving test data, leading to mismatches in training and test set distributions. As such, the problem of *domain adaptation* is of particular interest to the natural language processing community in order to build models which are robust this shift in distribution. For example, a model may be trained to predict the sentiment of product

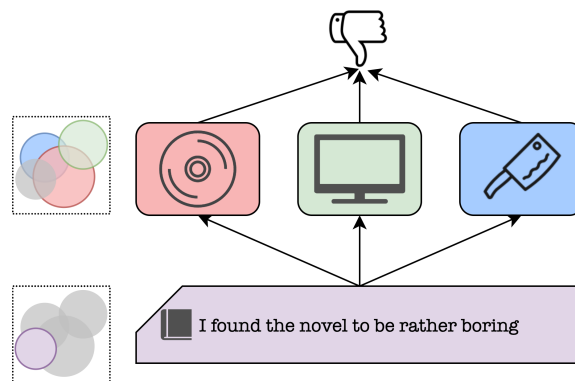


Figure 1: In multi-source domain adaptation, a model is trained on data drawn from multiple parts of the underlying distribution. At test time, the model must make predictions on data from a potentially non-overlapping part of the distribution.

reviews for DVDs, electronics, and kitchen goods, and must utilize this learned knowledge to predict the sentiment of a review about a book (Figure 1). This paper is concerned with this setting, namely *unsupervised multi-source domain adaptation*.

Multi-source domain adaptation is a well studied problem in deep learning for natural language processing. Prominent techniques are generally based on data selection strategies and representation learning. For example, a popular representation learning method is to induce domain invariant representations using unsupervised target data and domain adversarial learning (Ganin and Lempitsky, 2015). Adding to this, mixture of experts techniques attempt to learn both domain specific and global shared representations and combine their predictions (Guo et al., 2018; Li et al., 2018; Ma et al., 2019). These methods have been primarily studied using convolutional nets (CNNs) and recurrent nets (RNNs) trained from scratch, while the NLP community has recently begun to rely more and more on large pretrained transformer (LPX) models e.g. BERT (Devlin et al., 2019). To date

there has been some preliminary investigation of how LPX models perform under domain shift in the single source-single target setting (Ma et al., 2019; Han and Eisenstein, 2019; Rietzler et al., 2020; Gururangan et al., 2020). What is lacking is a study into the effects of and best ways to apply classic multi-source domain adaptation techniques with LPX models, which can give insight into possible avenues for improved application of these models in settings where there is domain shift.

Given this, we present a study into unsupervised multi-source domain adaptation techniques for large pretrained transformer models. Our main research question is: do mixture of experts and domain adversarial training offer any benefit when using LPX models? The answer to this is not immediately obvious, as such models have been shown to generalize quite well across domains and tasks while still learning representations which are not domain invariant. Therefore, we experiment with four mixture of experts models, including one novel technique based on attending to different domain experts; as well as domain adversarial training with gradient reversal. Surprisingly, we find that, while domain adversarial training helps the model learn more domain invariant representations, this does not always result in increased target task performance. When using mixture of experts, we see significant gains on out of domain rumour detection, and some gains on out of domain sentiment analysis. Further analysis reveals that the classifiers learned by domain expert models are highly homogeneous, making it challenging to learn a better mixing function than simple averaging.

## 2 Related Work

Our primary focus is multi-source domain adaptation with LPX models. We first review domain adaptation in general, followed by studies into domain adaptation with LPX models.

### 2.1 Domain Adaptation

Domain adaptation approaches generally fall into three categories: *supervised* approaches (e.g. Daumé (2007); Finkel and Manning (2009); Kulis et al. (2011)), where both labels for the source and the target domain are available; *semi-supervised* approaches (e.g. Donahue et al. (2013); Yao et al. (2015)), where labels for the source and a small set of labels for the target domain are provided; and lastly *unsupervised* approaches (e.g. Blitzer et al.

(2006); Ganin and Lempitsky (2015); Sun et al. (2016); Lipton et al. (2018)), where only labels for the source domain are given. Since the focus of this paper is the latter, we restrict our discussion to unsupervised approaches. A more complete recent review of unsupervised domain adaptation approaches is given in Kouw and Loog (2019).

A popular approach to unsupervised domain adaptation is to induce representations which are invariant to the shift in distribution between source and target data. For deep networks, this can be accomplished via domain adversarial training using a simple gradient reversal trick (Ganin and Lempitsky, 2015). This has been shown to work in the multi-source domain adaptation setting too (Li et al., 2018). Other popular representation learning methods include minimizing the covariance between source and target features (Sun et al., 2016) and using maximum-mean discrepancy between the marginal distribution of source and target features as an adversarial objective (Guo et al., 2018).

Mixture of experts has also been shown to be effective for multi-source domain adaptation. Kim et al. (2017) use attention to combine the predictions of domain experts. Guo et al. (2018) propose learning a mixture of experts using a point to set metric, which combines the posteriors of models trained on individual domains. Our work attempts to build on this to study how multi-source domain adaptation can be improved with LPX models.

### 2.2 Transformer Based Domain Adaptation

There are a handful of studies which investigate how LPX models can be improved in the presence of domain shift. These methods tend to focus on the data and training objectives for single-source single-target unsupervised domain adaptation. The work of Ma et al. (2019) shows that curriculum learning based on the similarity of target data to source data improves the performance of BERT on out of domain natural language inference. Additionally, Han and Eisenstein (2019) demonstrate that domain adaptive fine-tuning with the masked language modeling objective of BERT leads to improved performance on domain adaptation for sequence labelling. Rietzler et al. (2020) offer similar evidence for task adaptive fine-tuning on aspect based sentiment analysis. Gururangan et al. (2020) take this further, showing that significant gains in performance are yielded when progressively fine-tuning on in domain data, followed by task data,

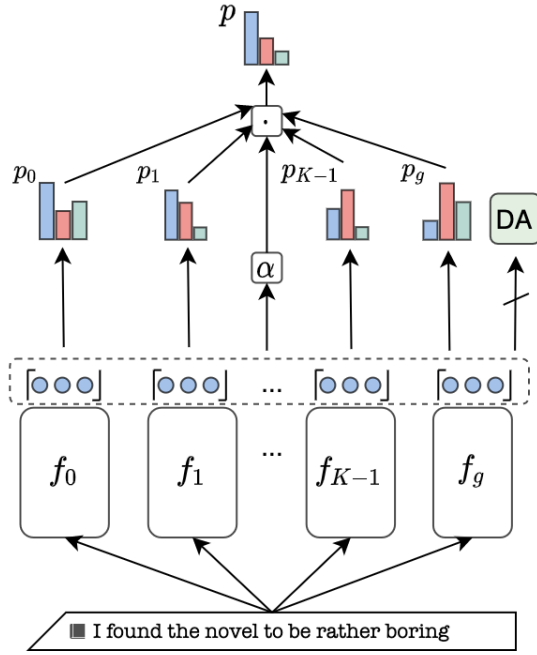


Figure 2: The overall approach tested in this work. A sample is input to a set of expert and one shared LPX model as described in §3.1. The output probabilities of these models are then combined using an attention parameter alpha (§3.1.1, §3.1.2, §3.1.3, §3.1.4). In addition, a global model  $f_g$  learns domain invariant representations via a classifier DA with gradient reversal (indicated by the slash, see §3.2).

using the masked language modeling objective of RoBERTa. Finally, Lin et al. (2020) explore whether domain adversarial training with BERT would improve performance for clinical negation detection, finding that the best performing method is a plain BERT model, giving some evidence that perhaps well-studied domain adaptation methods may not be applicable to LPX models.

What has not been studied, to the best of our knowledge, is the impact of domain adversarial training via gradient reversal on LPX models on natural language processing tasks, as well as if mixture of experts techniques can be beneficial. As these methods have historically benefited deep models for domain adaptation, we explore their effect when applied to LPX models in this work.

### 3 Methods

This work is motivated by previous research on domain adversarial training and mixture of domain experts for domain adaptation. In this, the data consists of  $K$  source domains  $\mathcal{S}$  and a target domain  $\mathcal{T}$ . The source domains consist of labelled datasets

$D_s, s \in \{1, \dots, K\}$  and the target domain consists only of unlabelled data  $U_t$ . The goal is to learn a classifier  $f$ , which generalizes well to  $\mathcal{T}$  using only the labelled data from  $\mathcal{S}$  and optionally unlabelled data from  $\mathcal{T}$ . We consider a base network  $f_z, z \in \mathcal{S} \cup \{g\}$  corresponding to either a domain specific network or a global shared network. These  $f_z$  networks are initialized using LPX models, in particular DistilBert (Sanh et al., 2019).

#### 3.1 Mixture of Experts Techniques

We study four different mixture of expert techniques: simple averaging, fine-tuned averaging, attention with a domain classifier, and a novel sample-wise attention mechanism based on transformer attention (Vaswani et al., 2017). Prior work reports that utilizing mixtures of domain experts and shared classifiers leads to improved performance when having access to multiple source domains (Guo et al., 2018; Li et al., 2018). Given this, we investigate if mixture of experts can have any benefit when using LPX models.

Formally, for a setting with  $K$  domains, we have set of  $K$  different LPX models  $f_k, k \in \{0 \dots K-1\}$  corresponding to each domain. There is also an additional LPX model  $f_g$  corresponding to a global shared model. The output predictions of these models are  $p_k, k \in \{0 \dots K-1\}$  and  $p_g$ , respectively. Since the problems we are concerned with are binary classification, these are single values in the range  $(0, 1)$ . The final output probability is calculated as a weighted combination of a set of domain expert probabilities  $\bar{\mathcal{K}} \subseteq \mathcal{S}$  and the probability from the global shared model. Four methods are used for calculating the weighting.

##### 3.1.1 Averaging

The first method is a simple averaging of the predictions of domain specific and shared classifiers. The final output of the model is

$$p_A(x, \bar{\mathcal{K}}) = \frac{1}{|\bar{\mathcal{K}}|+1} \sum_{k \in \bar{\mathcal{K}}} p_k(x) + p_g(x) \quad (1)$$

##### 3.1.2 Fine Tuned Averaging

As an extension to simple averaging, we fine tune the weight given to each of the domain experts and global shared model. This is performed via randomized grid search evaluated on validation data, after the models have been trained. A random integer between zero and ten is generated for each of the models, which is then normalized to a set

of probabilities  $\alpha_F$ . The final output probability is then given as follows.

$$p_F(x) = \sum_{k \in \bar{\mathcal{K}}} p_k(x) * \alpha_F^{(k)}(x) + p_g(x) * \alpha_F^{(g)}(x) \quad (2)$$

### 3.1.3 Domain Classifier

It was recently shown that curriculum learning using a domain classifier can lead to improved performance for single-source domain adaptation (Ma et al., 2019) when using LPX models. Inspired by this, we experiment with using a domain classifier as a way to attend to the predictions of domain expert models. First, a domain classifier  $f_C$  is trained to predict the domain of an input sample  $x$  given  $\mathbf{r}_g \in \mathbb{R}^d$ , the representation of the [CLS] token at the output of a LPX model. From the classifier, a vector  $\alpha_C$  is produced with the probabilities that a sample belongs to each source domain.

$$\alpha_C = f_C(x) = \text{softmax}(\mathbf{W}_C \mathbf{r}_g + b_C) \quad (3)$$

where  $\mathbf{W}_C \in \mathbb{R}^{d \times K}$  and  $b_C \in \mathbb{R}^K$ . The domain classifier is trained before the end-task network and is held static throughout training on the end-task. For this, a set of domain experts  $f_k$  are trained and their predictions combined through a weighted sum of the attention vector  $\alpha_C$ .

$$p_C(x) = \sum_{k \in S} p_k(x) * \alpha_C^{(k)}(x) \quad (4)$$

where the superscript  $(k)$  indexes into the  $\alpha_C$  vector. Note that in this case we only use domain experts and not a global shared model. In addition, the probability is always calculated with respect to each source domain.

### 3.1.4 Attention Model

Finally, a novel parameterized attention model is learned which attends to different domains based on the input sample. The attention method is based on the scaled dot product attention applied in transformer models (Vaswani et al., 2017), where a global shared model acts as a query network attending to each of the expert and shared models. As such, a shared model  $f_g$  produces a vector  $\mathbf{r}_g \in \mathbb{R}^d$ , and each domain expert produces a vector  $\mathbf{r}_k \in \mathbb{R}^d$ . First, for an input sample  $x$ , a probability for the end task is obtained from the classifier of each model yielding probabilities  $p_g$

and  $p_k, k \in 0 \dots K - 1$ . An attention vector  $\alpha_X$  is then obtained via the following transformations.

$$\mathbf{q} = \mathbf{g} \mathbf{Q}^T \quad (5)$$

$$\mathbf{k} = \begin{bmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_K \\ \mathbf{r}_g \end{bmatrix} \mathbf{K}^T \quad (6)$$

$$\alpha_X = \text{softmax}(\mathbf{q} \mathbf{k}^T) \quad (7)$$

where  $\mathbf{Q} \in \mathbb{R}^{d \times d}$  and  $\mathbf{K} \in \mathbb{R}^{d \times d}$ . The attention vector  $\alpha_X$  then attends to the individual predictions of each domain expert and the global shared model.

$$p_X(x, \bar{\mathcal{K}}) = \sum_{k \in \bar{\mathcal{K}}} p_k(x) * \alpha_X^{(k)}(x) + p_g(x) * \alpha_X^{(g)}(x) \quad (8)$$

To ensure that each model is trained as a domain specific expert, a similar training procedure to that of Guo et al. 2018 is utilized, described in §3.3.

## 3.2 Domain Adversarial Training

The method of domain adversarial adaptation we investigate here is the well-studied technique described in Ganin and Lempitsky (2015). It has been shown to benefit both convolutional nets and recurrent nets on NLP problems (Li et al., 2018; Gui et al., 2017), so is a prime candidate to study in the context of LPX models. Additionally, some preliminary evidence indicates that adversarial training might improve LPX generalizability for single-source domain adaptation (Ma et al., 2019).

To learn domain invariant representations, we train a model such that the learned representations maximally confuse a domain classifier  $f_d$ . This is accomplished through a min-max objective between the domain classifier parameters  $\theta_D$  and the parameters  $\theta_G$  of an encoder  $f_g$ . The objective can then be described as follows.

$$\mathcal{L}_D = \max_{\theta_D} \min_{\theta_G} -d \log f_d(f_g(x)) \quad (9)$$

where  $d$  is the domain of input sample  $x$ . The effect of this is to improve the ability of the classifier to determine the domain of an instance, while encouraging the model to generate maximally confusing representations via minimizing the negative loss. In practice, this is accomplished by training the model using standard cross entropy loss, but reversing the gradients of the loss with respect to the model parameters  $\theta_G$ .



### 3.3 Training

Our training procedure follows a multi-task learning setup in which the data from a single batch comes from a single domain. Domains are thus shuffled on each round of training and the model is optimized for a particular domain on each batch.

For the attention based (§3.1.4) and averaging (§3.1.1) models we adopt a similar training algorithm to Guo et al. (2018). For each batch of training, a meta-target  $t$  is selected from among the source domains, with the rest of the domains treated as meta-sources  $S' \in \mathcal{S} \setminus \{t\}$ . Two losses are then calculated. The first is with respect to all of the meta-sources, where the attention vector is calculated for only those domains. For target labels  $y_i$  and a batch of size  $N$  with samples from a single domain, this is given as follows.

$$\mathcal{L}_s = -\frac{1}{N} \sum_i y_i \log p_X(x, S') \quad (10)$$

The same procedure is followed for the averaging model  $p_A$ . The purpose is to encourage the model to learn attention vectors for out of domain data, thus why the meta-target is excluded from the calculation.

The second loss is with respect to the meta-target, where the cross-entropy loss is calculated directly for the domain expert network of the meta-target.

$$\mathcal{L}_t = -\frac{1}{N} \sum_i y_i \log p_t(x) \quad (11)$$

This allows each model to become a domain expert through strong supervision. The final loss of the network is a combination of the three losses described previously, with  $\lambda$  and  $\gamma$  hyperparameters controlling the weight of each loss.

$$\mathcal{L} = \lambda \mathcal{L}_s + (1 - \lambda) \mathcal{L}_t + \gamma \mathcal{L}_D \quad (12)$$

For the domain classifier (§3.1.3) and fine-tuned averaging (§3.1.2), the individual LPX models are optimized directly with no auxiliary mixture of experts objective. In addition, we experiment with training the simple averaging model directly.

## 4 Experiments and Results

We focus our experiments on text classification problems with data from multiple domains. To this end, we experiment with sentiment analysis from Amazon product reviews and rumour detection from tweets. For both tasks, we perform cross-validation on each domain, holding out a single

domain for testing and training on the remaining domains, allowing a comparison of each method on how well they perform under domain shift. The code to reproduce all of the experiments in this paper can be found here<sup>1</sup>.

**Sentiment Analysis Data** The data used for sentiment analysis come from the legacy Amazon Product Review dataset (Blitzer et al., 2007). This dataset consists of 8,000 total tweets from four product categories: books, DVDs, electronics, and kitchen and housewares. Each domain contains 1,000 positive and 1,000 negative reviews. In addition, each domain has associated unlabelled data. Following previous work we focus on the transductive setting (Guo et al., 2018; Ziser and Reichart, 2017) where we use the same 2,000 out of domain tweets as unlabelled data for training the domain adversarial models. This data has been well studied in the context of domain adaptation, making for easy comparison with previous work.

**Rumour Detection Data** The data used for rumour detection come from the PHEME dataset of rumourous tweets (Zubiaga et al., 2016). There are a total of 5,802 annotated tweets from 5 different events labelled as rumourous or non-rumourous (1,972 rumours, 3,830 non-rumours). Methods which have been shown to work well on this data include context-aware classifiers (Zubiaga et al., 2017) and positive-unlabelled learning (Wright and Augenstein, 2020). Again, we use this data in the transductive setting when testing domain adversarial training.

### 4.1 Baselines

**What’s in a Domain?** We use the model from Li et al. (2018) as a baseline for sentiment analysis. This model consists of a set of domain experts and one general CNN, and is trained with a domain adversarial auxiliary objective.

**Mixture of Experts** Additionally, we present the results from Guo et al. (2018) representing the most recent state of the art on the Amazon reviews dataset. Their method consists of domain expert classifiers trained on top of a shared encoder, with predictions being combined via a novel metric which considers the distance between the mean representations of target data and source data.

<sup>1</sup><https://github.com/copenlu/xformer-multi-source-domain-adaptation>

Method	Sentiment Analysis (Accuracy)					Rumour Detection (F1)					
	D	E	K	B	macroA	CH	F	GW	OS	S	$\mu$ F1
Li et al. 2018	77.9	80.9	80.9	77.1	79.2	-	-	-	-	-	-
Guo et al. 2018	87.7	89.5	90.5	87.9	88.9	-	-	-	-	-	-
Zubiaga et al. 2017	-	-	-	-	-	63.6	<b>46.5</b>	70.4	69.0	61.2	60.7
Basic	89.1	89.8	90.1	89.3	89.5	66.1	44.7	71.9	61.0	63.3	62.3
Adv-6	88.3	89.7	90.0	89.0	89.3	65.8	42.0	66.6	61.7	63.2	61.4
Adv-3	89.0	89.9	90.3	89.0	89.6	65.7	43.2	72.3	60.4	62.1	61.7
Independent-Avg	88.9	<b>90.6</b>	90.4	<b>90.0</b>	<b>90.0</b>	66.1	45.6	71.7	59.4	63.5	62.2
Independent-Ft	88.9	90.3	<b>90.8</b>	<b>90.0</b>	<b>90.0</b>	65.9	45.7	72.2	59.3	62.4	61.9
MoE-Avg	<b>89.3</b>	89.9	90.5	89.9	89.9	<b>67.9</b>	45.4	<b>74.5</b>	62.6	<b>64.7</b>	<b>64.1</b>
MoE-Att	88.6	90.0	90.4	89.6	89.6	65.9	42.3	72.5	61.2	63.3	62.2
MoE-Att-Adv-6	87.8	89.0	90.5	88.3	88.9	66.0	40.7	69.0	63.8	63.7	61.8
MoE-Att-Adv-3	88.6	89.1	90.4	88.9	89.2	65.6	42.7	73.4	60.9	61.0	61.8
MoE-DC	87.8	89.2	90.2	87.9	88.8	66.5	40.6	70.5	<b>70.8</b>	62.8	63.8

Table 1: Experiments for sentiment analysis in (D)VD, (E)lectronics, (K)itchen and housewares, and (B)ooks domains and rumour detection for different events ((C)harlie(H)ebdo, (F)erguson, (G)erman(W)ings, (O)ttawa(S)hooting, and (S)ydneySiege) using leave-one-out cross validation. Results are averaged across 5 random seeds. The results for sentiments analysis are in terms of accuracy and the results for rumour detection are in terms of F1.

**Zubiaga et al. 2017** Though not a domain adaptation technique, we include the results from Zubiaga et al. 2017 on rumour detection to show the current state of the art performance on this task. The model is a CRF, which utilizes a combination of content and social features acting on a timeline of tweets.

## 4.2 Model Variants

A variety of models are tested in this work. Therefore, each model is referred to by the following.

**Basic** Basic DistilBert with a single classification layer at the output.

**Adv- $X$**  DistilBert with domain adversarial supervision applied at the  $X$ 'th layer (§3.2).

**Independent-Avg** DistilBert mixture of experts averaged but trained individually (not with the algorithm described in §3.3).

**Independent-FT** DistilBert mixture of experts averaged with mixing attention fine tuned after training (§3.1.2), trained individually.

**MoE-Avg** DistilBert mixture of experts using averaging (§3.1.1).

**MoE-Att** DistilBert mixture of experts using our novel attention based technique (§3.1.4).

**MoE-Att-Adv- $X$**  DistilBert mixture of experts using attention and domain adversarial supervision applied at the  $X$ 'th layer.

**MoE-DC** DistilBert mixture of experts using a domain classifier for attention (§3.1.3).

## 4.3 Results

Our results are given in Table 1. Similar to the findings of Lin et al. (2020) on clinical negation, we see little overall difference in performance from both the individual model and the mixture of experts model when using domain adversarial training on sentiment analysis. For the base model, there is a slight improvement when domain adversarial supervision is applied at a lower layer of the model, but a drop when applied at a higher level. Additionally, mixture of experts provides some benefit, especially using the simpler methods such as averaging.

For rumour detection, again we see little performance change from using domain adversarial training, with a slight drop when supervision is applied at either layer. The mixture of experts methods overall perform better than single model methods, suggesting that mixing domain experts is still effective when using large pretrained transformer models. In this case, the best mixture of experts methods are simple averaging and static grid search for mixing weights, indicating the difficulty in learning an effective way to mix the predictions of domain experts. We elaborate on our findings further in §5. Additional experiments on domain adversarial training using Bert can be found in Table 2 in §A, where we similarly find that domain adversarial training leads to a drop in performance on both datasets.

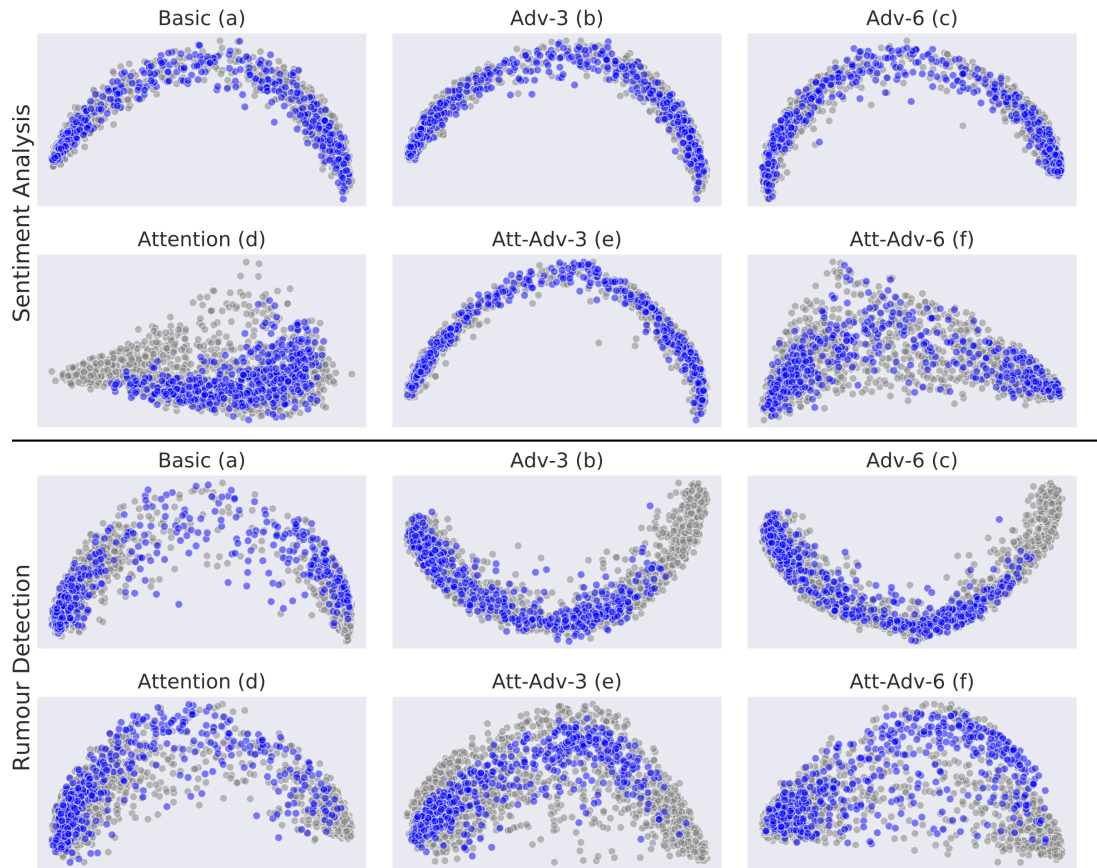


Figure 3: Final layer DistilBert embeddings for 500 randomly selected examples from each split for each tested model for sentiment data (top two rows) and rumour detection (bottom two rows). The blue points are out of domain data (in this case from Kitchen and Housewares for sentiment analysis and Sydney Siege for rumour detection) and the gray points are in domain data.

## 5 Discussion

We now discuss our initial research questions in light of the results we obtained, and provide explanations for the observed behavior.

### 5.1 What is the Effect of Domain Adversarial Training?

We present PCA plots of the representations learned by different models in Figure 3. These are the final layer representations of 500 randomly sampled points for each split of the data. In the ideal case, the representations for out of domain samples would be indistinguishable from the representations for in domain data.

In the case of basic DistilBert, we see a slight change in the learned representations of the domain adversarial models versus the basic model (Figure 3 top half, a-c) for sentiment analysis. When the attention based mixture of experts model is used, the representations of out of domain data cluster in one

region of the representation space (d). With the application of adversarial supervision, the model learns representations which are more domain agnostic. Supervision applied at layer 6 of DistilBert (plot f) yields a representation space similar to the version without domain adversarial supervision. Interestingly, the representation space of the model with supervision at layer 3 (plot e) yields representations similar to the basic classifier. This gives some potential explanation as to the similar performance of this model to the basic classifier on this split (kitchen and housewares). Overall, domain adversarial supervision has some effect on performance, leading to gains in both the basic classifier and the mixture of experts model for this split. Additionally, there are minor improvements overall for the basic case, and a minor drop in performance with the mixture of experts model.

The effect of domain adversarial training is more pronounced on the rumour detection data for the basic model (Figure 3 bottom half, a), where the rep-

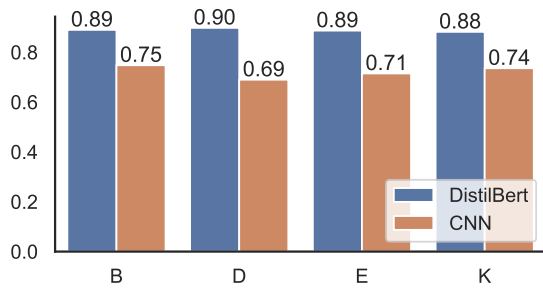


Figure 4: Comparison of agreement (Krippendorff’s alpha) between domain expert models when the models are either DistilBert or a CNN. Predictions are made on unseen test data by each domain expert, and agreement is measured between their predictions ((B)ooks, (D)VD, (E)lectronics, and (K)itchen). The overall agreement between the DistilBert experts is greater than the CNNs, suggesting that the learned classifiers are much more homogenous.

representations exhibit somewhat less variance when domain adversarial supervision is applied. Surprisingly, this leads to a slight drop in performance for the split of the data depicted here (Sydney Siege). For the attention based model, the variant without domain adversarial supervision (d) already learns a somewhat domain agnostic representation. The model with domain adversarial supervision at layer 6 (f) furthers this, and the classifier learned from these representations perform better on this split of the data. Ultimately, the best performing models for rumour detection do not use domain supervision, and the effect on performance on the individual splits are mixed, suggesting that domain adversarial supervision can potentially help, but not in all cases.

## 5.2 Is Mixture of Experts Useful with LPX Models?

We performed experiments with several variants of mixture of experts, finding that overall, it can help, but determining the optimal way to mix LPX domain experts remains challenging. Simple averaging of domain experts (§3.1.1) gives better performance on both sentiment analysis and rumour detection over the single model baseline. Learned attention (§3.1.4) has a net positive effect on performance for sentiment analysis and a negative effect for rumour detection compared to the single model baseline. Additionally, simple averaging of domain experts consistently outperforms a learned sample by sample attention. This highlights the difficulty

in utilizing large pretrained transformer models to learn to attend to the predictions of domain experts.

**Comparing agreement** To provide some potential explanation for why it is difficult to learn to attend to domain experts, we compare the agreement on the predictions of domain experts of one of our models based on DistilBert, versus a model based on CNNs (Figure 4). CNN models are chosen in order to compare the agreement using our approach with an approach which has been shown to work well with mixture of experts on this data (Guo et al., 2018). Each CNN consists of an embedding layer initialized with 300 dimensional FastText embeddings (Bojanowski et al., 2017), a series of 100 dimensional convolutional layers with widths 2, 4, and 5, and a classifier. The end performance is on par with previous work using CNNs (Li et al., 2018) (78.8 macro averaged accuracy, validation accuracies of the individual models are between 80.0 and 87.0). Agreement is measured using Krippendorff’s alpha (Krippendorff, 2011) between the predictions of domain experts on test data.

We observe that the agreement between DistilBert domain experts on test data is significantly higher than that of CNN domain experts, indicating that the learned classifiers of each expert are much more similar in the case of DistilBert. Therefore, it will potentially be more difficult for a mixing function on top of DistilBert domain experts to gain much beyond simple averaging, while with CNN domain experts, there is more to be gained from mixing their predictions. This effect may arise because each DistilBert model is highly pre-trained already, hence there is little change in the final representations, and therefore similar classifiers are learned between each domain expert.

## 6 Conclusion

In this work, we investigated the problem of multi-source domain adaptation with large pretrained transformer models. Both domain adversarial training and mixture of experts techniques were explored. While domain adversarial training could effectively induce more domain agnostic representations, it had a mixed effect on model performance. Additionally, we demonstrated that while techniques for mixing domain experts can lead to improved performance for both sentiment analysis and rumour detection, determining a beneficial mixing of such experts is challenging. The best method we tested was a simple averaging of



the domain experts, and we provided some evidence as to why this effect was observed. We find that LPX models may be better suited for data-driven techniques such as that of Gururangan et al. (2020), which focus on inducing a better prior into the model through pretraining, as opposed to techniques which focus on learning a better posterior with architectural enhancements. We hope that this work can help inform researchers of considerations to make when using LPX models in the presence of domain shift.

## Acknowledgements



This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801199.

## References

- John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, Boom-Boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447.
- John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. [Domain Adaptation with Structural Correspondence Learning](#). In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney, Australia. Association for Computational Linguistics.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Hal Daumé, III. 2007. [Frustratingly Easy Domain Adaptation](#). In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT 2019*, pages 4171–4186.
- Jeff Donahue, Judy Hoffman, Erik Rodner, Kate Saenko, and Trevor Darrell. 2013. [Semi-supervised Domain Adaptation with Instance Constraints](#). In *CVPR*, pages 668–675. IEEE Computer Society.
- Jenny Rose Finkel and Christopher D. Manning. 2009. [Hierarchical Bayesian Domain Adaptation](#). In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 602–610, Boulder, Colorado. Association for Computational Linguistics.
- Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised Domain Adaptation by Backpropagation. In *International Conference on Machine Learning*, pages 1180–1189.
- Tao Gui, Qi Zhang, Haoran Huang, Minlong Peng, and Xuan-Jing Huang. 2017. Part-of-Speech Tagging for Twitter with Adversarial Neural Networks. In *EMNLP 2017*, pages 2411–2420.
- Jiang Guo, Darsh Shah, and Regina Barzilay. 2018. Multi-Source Domain Adaptation with Mixture of Experts. In *EMNLP 2018*, pages 4694–4703.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks.
- Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised Domain Adaptation of Contextualized Embeddings: A Case Study in Early Modern English. pages 4229–4239.
- Young-Bum Kim, Karl Stratos, and Dongchan Kim. 2017. Domain Attention With an Ensemble of Experts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 643–653.
- Wouter Marco Kouw and Marco Loog. 2019. A Review of Domain Adaptation Without Target Labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Klaus Krippendorff. 2011. Computing Krippendorff’s Alpha-Reliability.
- Brian Kulis, Kate Saenko, and Trevor Darrell. 2011. [What You Saw is Not What You Get: Domain Adaptation Using Asymmetric Kernel Transforms](#). In *CVPR*, pages 1785–1792. IEEE Computer Society.
- Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. What’s in a Domain? Learning Domain-Robust Text Representations Using Adversarial Training. pages 474–479.
- Chen Lin, Steven Bethard, Dmitriy Dligach, Farig Sad-eque, Guergana Savova, and Timothy A Miller. 2020. Does BERT Need Domain Adaptation for Clinical Negation Detection? *Journal of the American Medical Informatics Association*, 27(4):584–591.
- Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. 2018. [Detecting and Correcting for Label Shift with Black Box Predictors](#). In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 3128–3136. PMLR.

- Xiaofei Ma, Peng Xu, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2019. Domain Adaptation with BERT-based Domain Classification and Data Selection. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 76–83.
- Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or Get Left Behind: Domain Adaptation Through Bert Language Model Finetuning for Aspect-Target Sentiment Classification. pages 4933–4941.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv preprint arXiv:1910.01108*.
- Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. [Return of Frustratingly Easy Domain Adaptation](#). In *AAAI*, pages 2058–2065. AAAI Press.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.
- Dustin Wright and Isabelle Augenstein. 2020. Claim Check-Worthiness Detection as Positive Unlabelled Learning. In *Findings of EMNLP*. Association for Computational Linguistics.
- Ting Yao, Yingwei Pan, Chong-Wah Ngo, Houqiang Li, and Tao Mei. 2015. [Semi-supervised Domain Adaptation with Subspace Learning for Visual Recognition](#). In *CVPR*, pages 2142–2150. IEEE Computer Society.
- Yftah Ziser and Roi Reichart. 2017. Neural Structural Correspondence Learning for Domain Adaptation. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 400–410.
- Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2017. Exploiting Context for Rumour Detection in Social Media. In *International Conference on Social Informatics*, pages 109–123. Springer.
- Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads. *PloS one*, 11(3).

## A BERT Domain Adversarial Training Results

Additional results on domain adversarial training with Bert can be found in [Table 2](#).

## B Reproducibility

### B.1 Computing Infrastructure

All experiments were run on a shared cluster. Requested jobs consisted of 16GB of RAM and 4 Intel Xeon Silver 4110 CPUs. We used a single NVIDIA Titan X GPU with 12GB of RAM.

### B.2 Average Runtimes

The average runtime performance of each model is given in [Table 3](#). Note that different runs may have been placed on different nodes within a shared cluster, thus why large time differences occurred.

### B.3 Number of Parameters per Model

The number of parameters in each model is given in [Table 4](#).

### B.4 Validation Performance

The validation performance of each tested model is given in [Table 5](#).

### B.5 Evaluation Metrics

The primary evaluation metrics used were accuracy and F1 score. For accuracy, we used our implementation provided with the code. The basic implementation is as follows.

$$\text{accuracy} = \frac{tp + tn}{tp + fp + tn + fn}$$

We used the sklearn implementation of precision\_recall\_fscore\_support for F1 score, which can be found here: [https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision\\_recall\\_fscore\\_support.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html). Briefly:

$$p = \frac{tp}{tp + fp}$$

$$r = \frac{tp}{tp + fn}$$

$$F1 = \frac{2 * p * r}{p + r}$$

where  $tp$  are true positives,  $fp$  are false positives, and  $fn$  are false negatives.

## B.6 Hyperparameters

We performed an initial hyperparameter search to obtain good hyperparameters that we used across models. The bounds for each hyperparameter was as follows:

- Learning rate: [0.00003, 0.00004, 0.00002, 0.00001, 0.00005, 0.0001, 0.001].
- Weight decay: [0.0, 0.1, 0.01, 0.005, 0.001, 0.0005, 0.0001].
- Epochs: [2, 3, 4, 5, 7, 10].
- Warmup steps: [0, 100, 200, 500, 1000, 5000, 10000].
- Gradient accumulation: [1,2]

We kept the batch size at 8 due to GPU memory constraints and used gradient accumulation instead. We performed a randomized hyperparameter search for 70 trials. Best hyperparameters are chosen based on validation set performance (accuracy for sentiment data, F1 for rumour detection data). The final hyperparameters selected are as follows:

- Learning rate:  $3e-5$ .
- Weight decay: 0.01.
- Epochs: 5.
- Warmup steps: 200.
- Batch Size: 8
- Gradient accumulation: 1

Additionally, we set the objective weighting parameters to  $\lambda = 0.5$  for the mixture of experts models and  $\gamma = 0.003$  for the adversarial models, in line with previous work ([Guo et al., 2018](#); [Li et al., 2018](#)).

## B.7 Links to data

- Amazon Product Reviews ([Blitzer et al., 2007](#)): <https://www.cs.jhu.edu/~mdredze/datasets/sentiment/>
- PHEME ([Zubiaga et al., 2016](#)): [https://figshare.com/articles/PHEME\\_dataset\\_for\\_Rumour\\_Detection\\_and\\_Veracity\\_Classification/6392078](https://figshare.com/articles/PHEME_dataset_for_Rumour_Detection_and_Veracity_Classification/6392078).

Method	Sentiment Analysis (Accuracy)					Rumour Detection (F1)					
	D	E	K	B	macroA	CH	F	GW	OS	S	$\mu$ F1
Bert	90.3	91.6	91.7	90.4	91.0	66.4	46.2	68.3	67.3	62.3	63.3
Bert-Adv-12	89.8	91.4	91.2	90.1	90.6	66.6	47.8	62.5	65.3	62.8	62.5
Bert-Adv-4	89.9	91.1	91.7	90.4	90.8	65.6	43.6	71.0	68.1	60.8	62.8

Table 2: Experiments for sentiment analysis in (D)VD, (E)lectronics, (K)itchen and housewares, and (B)ooks domains and rumour detection for different events ((C)harlie(H)ebdo, (F)erguson, (G)erman(W)ings, (O)ttawa(S)hooting, and (S)ydneySiege) using leave-one-out cross validation for BERT. Results are averaged across 3 random seeds. The results for sentiments analysis are in terms of accuracy and the results for rumour detection are in terms of F1.

Method	Sentiment Analysis	Rumour Detection
Basic	0h44m37s	0h23m52s
Adv-6	0h54m53s	0h59m31s
Adv-3	0h53m43s	0h57m29s
Independent-Avg	1h39m13s	1h19m27
Independent-Ft	1h58m55s	1h43m13
MoE-Avg	2h48m23s	4h03m46s
MoE-Att	2h49m44s	4h07m3s
MoE-Att-Adv-6	4h51m38s	4h58m33s
MoE-Att-Adv-3	4h50m13s	4h54m56s
MoE-DC	3h23m46s	4h09m51s

Table 3: Average runtimes for each model on each dataset (runtimes are taken for the entire run of an experiment).

Method	Sentiment Analysis	Rumour Detection
Basic	66,955,010	66,955,010
Adv-6	66,958,082	66,958,850
Adv-3	66,958,082	66,958,850
Independent-Avg	267,820,040	334,775,050
Independent-Ft	267,820,040	334,775,050
MoE-Avg	267,820,040	334,775,050
MoE-Att	268,999,688	335,954,698
MoE-Att-Adv-6	269,002,760	335,958,538
MoE-Att-Adv-3	269,002,760	335,958,538
MoE-DC	267,821,576	334,777,354

Table 4: Number of parameters in each model

Method	Sentiment Analysis (Acc)	Rumour Detection (F1)
Basic	91.7	82.4
Adv-6	91.5	83.3
Adv-3	91.2	83.4
Independent-Avg	92.7	82.8
Independent-Ft	92.6	82.5
MoE-Avg	92.2	83.5
MoE-Att	92.0	83.3
MoE-Att-Adv-6	91.2	83.3
MoE-Att-Adv-3	91.4	82.8
MoE-DC	89.8	84.6

Table 5: Average validation performance for each of the models on both datasets.