# Domain Divergences: a Survey and Empirical Analysis

**Abhinav Ramesh Kashyap**[†], **Devamanyu Hazarika**[†], **Min-Yen Kan**[†], **Roger Zimmermann** [†]
[†]National University of Singapore, Singapore
{abhinav, hazarika, kanmy, rogerz}@comp.nus.edu.sg



## Abstract

Domain divergence plays a significant role in estimating the performance of a model when applied to new domains. While there is significant literature on divergence measures, choosing an appropriate divergence measures remains difficult for researchers. We address this shortcoming by both surveying the literature and through an empirical study. We contribute a taxonomy of divergence measures consisting of three groups — Information-theoretic, Geometric, and Higher-order measures — and identify the relationships between them. We then ground the use of divergence measures in three different application groups – 1) Data Selection, 2) Learning Representation, and 3) Decisions in the Wild. From this, we identify that Information-theoretic measures are prevalent for 1) and 3), and higher-order measures are common for 2). To further help researchers, we validate these uses empirically through a correlation analysis of performance drops. We consider the current contextual word representations (CWR) to contrast with the older word distribution based representations for this analysis. We find that traditional measures over word distributions still serve as strong baselines, while higher-order measures with CWR are effective.

## 1 Introduction

Machine learning models perform poorly when they are tested on data that comes from a different target domain. Target domain performance largely depends on the divergence between the domains (Ben-David et al., 2010). Measuring domain divergence efficiently is important for adapting models in the new domain – the topic of *domain adaptation*. Domain adaptation also applies in the tasks of predicting the model performance drop in real-world settings (Van Asch and Daelemans, 2010), and in choosing among alternate models (Xia et al., 2020).

Research has invested much effort to define and measure domain divergence. Linguists use *register variation* to capture varieties in text – the difference in distribution of the prevalent features between two registers (Biber and Conrad, 2009). (Ben-David et al., 2010) introduce a probabilistic measure $\mathcal{H}$-divergence – to measure the difference between feature-distributions in the source and target domain. Further, information-theoretic measures like Leibler (KL) and Jenssen Shannon (JS) divergence based on surface features of text are also used for different applications (Plank and van Noord, 2011; Van Asch and Daelemans, 2010). In recent times, there is an emerging class of measures like Maximum Mean Discrepancy (MMD) and central Moment Discrepancy (CMD) (Gretton et al., 2007; Zellinger et al., 2017) that consider higher order moments of random variables. These measures are utilised for different applications in NLP, albeit in silos.

Given a plethora of divergence measures in the NLP literature, researchers are not clear on which measure is suitable for a given NLP task. To aid them, we first comprehensively review the NLP literature on domain divergences. Unlike, prior surveys in NLP, which focus on domain adaptation on particular tasks like, machine translation (Chu and Wang, 2018), and statistical (non-neural network) models (Jiang, 2007; Margolis, 2011), our work takes a different tack. We study domain adaptation through the vehicle of *domain divergence measures*. We group divergence measures into a taxonomy of three classes: Information Theoretic, Geometric, and Higher-Order measures. We then identify relationships between measures under each class. To identify the common class of measures used in different NLP applications, we recognise three novel applications of divergence — Data Selection, Learning Representations, and Decisions in the Wild — and organise their litera-

ture. We find that Information Theoretic measures over word distributions are common for Data Selection and Decisions in the wild, while Higher-order measures over continuous features are common for Learning representations.

As divergence between domains is a major determiner of target domain performance, a good domain divergence metric ideally predicts the corresponding performance drop of a model when applied on a new domain. We further help researchers identify appropriate measures, by performing a correlation analysis over 130 domain adaptation scenarios and three standard and varied NLP tasks: Part of Speech Tagging (POS), Named Entity Recognition (NER), and Sentiment Analysis. While information theoretic measures over word distributions are popular in the literature, are higher order measures calculated over contextual word representations better indicators of performance drop? We find that while higher order measures are better, traditional measures are still reliable indicators of performance drop.

We limit our survey to works that have a focus on domain divergence measures, and which consider unsupervised domain adaptation (UDA); i.e., where there is no annotated data available in the target domain, which is thus more practical yet more challenging. For a complete treatment of neural networks and UDA in NLP, we refer the reader to (Ramponi and Plank, 2020). Also, we do not treat multilingual work — although cross lingual transfer may be considered as an extreme form of domain adaptation, measuring distance between languages requires different divergence measures, outside of our purview.

## 2 A Taxonomy of Domain Divergence Measures

We devise a taxonomy for domain divergence measures, shown in Figure 1. Our taxonomy contains three main class of measures. Individual measures belong to a single category, where relationships can exist between measures from different categories. We provide detailed description of individual measures in Appendix A.

**Geometric measures** calculate the distance between two vectors in a metric space. As a domain divergence measure, they are used to calculate the distance between features of instances ($tf.idf$, continuous representations, etc.) from different domains. The P-norm is a generic form of the distance between two vectors, where Manhattan distance *(p=1)* and Euclidean distance *(p=2)* are common settings. Cosine (Cos) uses the cosine of the angle between two vectors to measure similarity, with $1 - Cos$ measuring the distance. Geometric measures are easy to calculate, but are ineffective for high-dimensional vectors.

**Information-theoretic measures** captures the distance between probability distributions. For example, cross entropy over n-gram word distributions are extensively used to rank sentences in a domain for further selection. $f$-divergence (Csiszár, 1972) are a general family of divergences where $f$ is a convex function. Different formulations of the $f$ function lead separately to KL and JS divergence. Chen and Cardie (2018) show that reducing $f$-divergence measure is equivalent to reducing the PAD measures (see next section). Another special case of $f$-divergence are the family of $\alpha$ divergences, themselves generalisations of KL-Div. Renyi Divergence is a member of the $\alpha$-divergences and tends towards KL-Div as as $\alpha \rightarrow 1$ (Edge Ⓐ); Often applied to optimal transport problems, Wasserstein distance measures the distance as the amount of work needed to convert one probability distribution to the other and finds applications in general machine learning and domain adaptation specifically. There are a variety of information theory based measures and they are linked to each other. KL-Div is also related to Cross Entropy (CE). In this paper, CE refers to other measures based on entropy.

**Higher-Order** measures consider matching higher order moments of random variables. Their properties are amenable to end-to-end learning based domain adaptation and has been adopted extensively recently. Maximum Mean Discrepancy (MMD) is one such measure which considers matching first order moments of variables in a Reproducible Kernel Hilbert Space. On the other hand, CORAL (Sun et al., 2017) considers second order moments and CMD (Zellinger et al., 2017) consider higher order moments. CORAL and CMD are desirable because they avoid computationally expensive kernel matrix computations. KL-Div can also be considered as matching the first-order moment (Zellinger et al., 2017); Edge Ⓑ.

A few other measures do not have ample support in the literature. These include information-theoretic measures such as Bhattacharya coeffi-

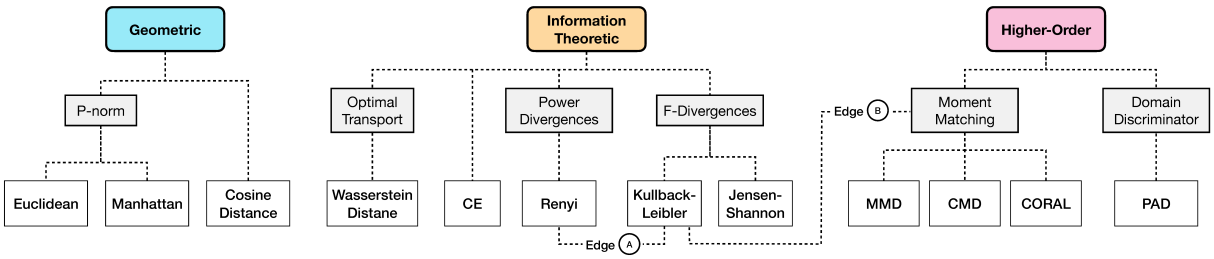



Figure 1: Taxonomy for divergence measures. i) **Geometric** measures the distance between vectors in a metric space ii) **Information- theoretic** measures the distance between probability distributions and iii) **Higher-order** measures the distance between distributions considering higher moments or the distance between representations or their projections in a nonlinear space.

cient, higher-order measures like PAD* (Elsahar and Gallé, 2019), Word Vector Variance (WVV), and Term Vocabulary Overlap (TVO) (Dai et al., 2019).

## 3 Applications of Domain Divergences

Our key observation of the literature is that there are three primary families of applications for divergence measures in NLP (cf. Table 2): $(i)$ **Data Selection**: selects a subset of text from a source domain that shares similar characteristics as target domain. The selected subset is then used to learn a target domain model. $(ii)$ **Learning Representations**: to align source and target domain feature distributions to ensure domain-invariance . $(iii)$ **Decisions in the Wild**: helps practitioners predict the performance or drops in performance on new data which can be subsequently used to drive decisions about annotating more data, choosing alternate models etc. Our taxonomy synthesises the diversity and the prevalence of the divergence measures in NLP.

### 3.1 Data Selection

Divergence measures are used to select a subset of text from the source domain that shares similar characteristics to the target domain. The selected data serves as supervised data for training models in the target domain. They are also used for learning self-supervised language modeling representations.

Simple word-level and surface-level text features like word frequency distributions, $tf.idf$ weighted distributions have sufficient power to distinguish between text varieties and help in data selection. Geometric measures like cosine, used with word frequency distributions are effective for selecting data in parsing and POS tagging (Plank and van Noord, 2011). In another work, Remus (2012) show that that JS-Div – an information theoretic measure, is effective for sentiment analysis. While these features are useful to select supervised data for an end-task, they also can be used to select data to pre-train language-models subsequently used for NER. (Dai et al., 2019) use Term Vocabulary Overlap for selecting data for pretraining language models. Geometric and Information-theoretic measures with word level distributions are inexpensive to calculate. However, estimating the distributions reliably needs large-scale data.

Continuous or distributed representations of words alleviate the short-comings of frequency-based probability distribution of words and are useful for data selection. Representations like Continuous Bag of Words (CBOW) and Skip-gram vectors (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), project words which have similar context closer together in space. However, they are static and do not change according to the context in which it is used. Contextual word representations produce different embedding depending on the context. Such representations which are mostly from neural networks (Devlin et al., 2019; Peters et al., 2018) help in capturing contextual similarities between words in two different domains. A Geometric measure – Word Vector Variance along with continuous word representations is useful for selecting data – similar in tenor to target data for pretraining neural networks (Dai et al., 2019). Further, P-norm and representations from pretrained neural machine translation models have been found effective for machine translation (Wang et al., 2017). Recently, (Aharoni and Goldberg, 2020) showed that contextual representation from top layers of BERT cluster according to different domains and can be used to perform data selection for neural machine translation (Aharoni and Goldberg, 2020).

Language models determine the probability of a sentence. If a language model trained on the target domain assigns high probability to a sentence from the source domain, then the sentence should have similar characteristics to the source domain. Cross Entropy and their variants capture this notion of similarity between two domains. They have been extensively used for data selection in statistical machine translation (Yasuda et al., 2008; Moore and Lewis, 2010; Axelrod et al., 2011; Duh et al., 2013; Liu et al., 2014). However, cross entropy based methods for data selection does not work effectively for neural machine translation (NMT). (van der Wees et al., 2017; Silva et al., 2018). Instead, (van der Wees et al., 2017) come up with a dynamic subset selection where new subset is chosen every epoch during NMT training.

Similar to language model, probabilistic scores from supervised classifiers which distinguish between samples from two domains can help in data selection. The probabilities assigned by such discriminators in construing source domain text as target domain text serves as the divergence measure for data selection in machine translation (Chen and Huang, 2016). However, they would require considerable amount of target domain data which is not available always. Alternatively, instead of training domain discriminators and then using it for data selection, (Chen et al., 2017) train a discriminator and selector in an alternating optimisation manner.

**From the literature review we find that different measures are found to be effective for different NLP tasks.** (Ruder and Plank, 2017) argue that owing to the different characteristics of the task, different methods can be useful. Hence, instead of using measures individually, they show that learning a linear combination of different measure to be useful for NER, parsing and sentiment analysis. However, this is not always possible, especially in unsupervised domain adaptation where there is no supervised data in target domain. **From Table 2, we note that information theoretic measures and geometric measures based on frequency based distributions and continuous representations are common for text prediction and structured prediction tasks.** The effectiveness of higher order measures are still not ascertained for these tasks.

Further, we find that for SMT data selection, variants of Cross Entropy measures find extensive use cases. However, the conclusions of (van der Wees et al., 2017) are more measured regarding the benefits of CE measures and the liked for NMT. **Contextual word representations with cosine similarity has found some initial exploration for neural machine translation with higher order measures yet to be explored for data selection in Neural Machine Translation**. We note that, the literature pays closer attention to data selection for machine translation compared to other tasks, owing to its popularity and practical applications. Also, given thousands of languages, obtaining parallel sentences between each combination of language is impractical.

### 3.2 Learning Representations

Domain adaptation aims to learn a model that works across different domains. One way to achieve this is to learn representations that are domain-invariant while being discriminative to perform well on a task (Ganin et al., 2015; Ganin and Lempitsky, 2015). Here, we limit our review to works utilising divergence measures. We exclude feature-based UDA methods like Structural Corresponding Learning (SCL) (Blitzer et al., 2006), Autoencoder-SCL, pivot based language models (Ziser and Reichart, 2017, 2018, 2019; Ben-David et al., 2020).

The theory of domain divergence (Ben-David et al., 2010) shows that the target domain error is bounded by the source domain error and domain divergence ($\mathcal{H}$-divergence). PAD, an approximation of $\mathcal{H}$-divergence, is large when a domain discriminator's error is small. Here, the discriminator is a supervised model that distinguishes samples between source and target domains. For domain-invariance, learned representations should not be able to distinguish source and target domain samples.

Motivated by $\mathcal{H}$ divergence, *Domain Adversarial Neural Networks* (DANN) (Ganin et al., 2015) aim to learn domain-invariant representations by using a domain discriminator. The network employs a min–max game — between the representation learner and the domain discriminator — inspired by Generative Adversarial Networks (Goodfellow et al., 2014). The encoder is trained by reversing the gradients calculated for the discriminator. In a later work, Bousmalis et al. (2016) argue that domain-specific peculiarities are lost in a DANN and propose *Domain Separation Networks* (DSN), where both domain-specific and -

invariant representations are captured in a *shared–private* network. DSN is flexible in its choice of divergence measures and finds PAD to perform better over MMD.

Obtaining domain invariant representations, is desirable for many different NLP tasks, especially for tasks like sequence labelling where annotating large amounts of data is hard. They are typically used when there is a single source domain and a single target domain – for sentiment analysis (Ganin et al., 2016), NER (Zhou et al., 2019), stance detection (Xu et al., 2019), machine translation (Britz et al., 2017; Zeng et al., 2018). The direct application of DANN and DSN to a variety of tasks, is a proof of their generality.

DANN and DSN have been inventively applied in other situations. Text from two different periods of time can be considered as two different domains and (Kim et al., 2017) perform intent classification on such text. In another work, (Gui et al., 2017) consider clean formal news-wire data as source domain and noisy, colloquial, unlabeled twitter data as the target domain and use adversarial learning to learn robust representations for POS. Common sense knowledge graph can help in learning domain invariant representations as well. (Ghosal et al., 2020) condition DANN with external common sense knowledge graph using graph convolutional neural networks for sentiment analysis. In contrast to the above works, (Wang et al., 2018) use MMD outside the adversarial learning framework. They use MMD to learn to reduce the discrepancy between neural network representations belonging to two different domains. Such concepts has been explored in computer vision by (Tzeng et al., 2014).

Complimentary information can be available in multiple domains that can help performance in a target domain. DANN and DSN networks have been extended to such multi-domain domain adaptation. (Ding et al., 2019) perform multi-domain intent classification. (Chen and Cardie, 2018; Li et al., 2018; Guo et al., 2018; Wright and Augenstein, 2020) perform sentiment analysis from multiple sources and single target domain using domain adversarial concepts. DSN has also been used for multi-source domain adaptation in machine translation (Gu et al., 2019; Wang et al., 2019).

Multi-task learning helps in improving generalisation by modeling two complimentary tasks. A key to obtaining benefits is to learn a shared representation that captures the common features of two tasks. However, such representations might still contain task-specific peculiarities. The shared-private model of DSN can help in disentangling such representations and has been used for sentiment analysis (Liu et al., 2017) and Chinese specific NER and word segmentation (Cao et al., 2018).

Although we do not deal with multi-lingual learning in this work, we have to note that, DANN and DSN can be extended to learn language agnostic representations useful for text classification and structured prediction works (Chen et al., 2018; Zou et al., 2018; Yasunaga et al., 2018).

Most of the works that adopt DANN and DSN framework reduce either the PAD or MMD distance between distributions. However, reducing the divergences, combined with other auxiliary task specific loss functions can result in training instabilities and vanishing gradient problem when the domain discriminator becomes increasingly more accurate (Shen et al., 2018). Using other higher order measures can result in better, stable learning. CMD has been used for sentiment analysis (Zellinger et al., 2017; Peng et al., 2018). Wasserstein distance has been used for duplicate question detection (Shah et al., 2018) and to learn domain-invariant attention distributions for emotional regression (Zhu et al., 2019).

**We can see from Table 2 that, most works extend the popular DSN framework to learn domain invariant representations, in different scenarios across NLP-tasks.** The original work from (Bousmalis et al., 2016) includes MMD divergence besides PAD, which is not adopted in subsequent works, possibly due to the reported poor performance. Most of the works require careful balancing between multiple objective functions (Han and Eisenstein, 2019), which can affect the stability of training. **The stability of training can be improved by selecting appropriate divergence measures like CMD (Zellinger et al., 2017) and Wasserstein Distance (Arjovsky et al., 2017)** and we envision more working using such measures owing to their advantages.

### 3.3 Decisions in the Wild

Models perform poorly when they are deployed in the real world. The performance degrades due to the difference in distribution between training and test data. Such performance degradation, can

always be alleviated by expensive large-scale annotation in the new domain. However, they are expensive and given thousands of domains – becomes infeasible. Thus predicting the performance in a new domain – where there is no labelled data is important. In recent times, it has received many theoretical considerations (Rosenfeld et al., 2020; Chuang et al., 2020; Steinhardt, 2016). As many researchers and engineers deploy models in the real world, it is important from a practical perspective. Empirically, works in NLP consider the divergence that exists between data in an unknown domain and known dataset to measure drops in performance.

Simple measures based on word level features has been used to predict the performance of a machine learning model in a new domain. Information theoretic measures like Renyi-Div, KL-Div has been used for predicting performance drops in POS (Van Asch and Daelemans, 2010) and Cross-Entropy based measure has been used for dependency parsing (Ravi et al., 2008). Prediction of performance can also be useful for machine translation where obtaining parallel data is hard. Based on distance features between languages and dataset features, (Xia et al., 2020) predict performance of the model on new languages for a variety of NLP tasks, including machine translation, POS, parsing etc. Such performance prediction models have also been done in the past for statistical machine translation (Birch et al., 2008; Specia et al., 2013).

However, (Ponomareva and Thelwall, 2012) argue that predicting *drop in performance* is more appropriate (why?) compared to just performance. They find that JS-Div is effective for predicting drop in performance of Sentiment Analysis systems. Only recently, predicting model failures in practical deployments from an empirical viewpoint has regained attention. (Elsahar and Gallé, 2019) find the efficacy of test higher-order measures to predict the drop in performance for POS and SA. **However, analysing performance drops using contextual word representations is still lacking**. We tackle this in the next section.

## 4 Empirical Analysis

How relevant are traditional measures over word distributions for measuring domain divergences? We examine this question given that contextual word representations such as BERT, ELMo, DistilBERT (Devlin et al., 2019; Peters et al., 2018; Sanh et al., 2019) are widespread and given that higher-order measures are increasingly being used to learn representations.

We perform an empirical study to assess their suitability for three important NLP tasks. POS, NER, and SA. We leave Natural language generation and MT for future work.

Performance difference between the source and the target domain depends on the divergence between feature distributions between domain (Ben-David et al., 2010). Like many other works (Ganin et al., 2016) we assume a co-variate shift, where the marginal distribution over features change, but the conditional label distributions does not– that is, $P_{\mathcal{D}_s}(y|x) = P_{\mathcal{D}_T}(y|x)$ $\mathcal{P}_{\mathcal{D}_s}(x) \neq P_{\mathcal{D}_T}(x)$. Although, difference in conditional label distribution can increase the H-Divergence measure (Wisniewski and Yvon, 2019), it requires labels in the target domain for assessment. In this work we assume no labelled data in the target domain, like in a realistic setting.

For all our experiments, unless otherwise mentioned, we use the DistilBERT (Sanh et al., 2019) pre-trained transformer model. It has competitive performance to BERT, but has faster inference times and lower resource requirements. We leave experimentation with other BERT variants, such as Roberta (Liu et al., 2019), for future work. For every text segment, we obtain the activations from the final layer and average-pool the representations. For domain adaptation scenarios, we train the models on the source domain training split. We test the best model from the grid search on the test dataset of the same domain and also the test dataset of the other domains (*cf.* Appendix C).

**Datasets:** For POS, we select 5 different corpora from the English Word Tree Bank of Universal Dependency corpus (Nivre et al., 2016)[1] and also include the GUM, Lines, and ParTUT datasets. We follow Elsahar and Gallé (2019) and consider these as 8 domains. For NER, we consider CONLL 2003 (Tjong Kim Sang and De Meulder, 2003), Emerging and Rare Entity Recognition Twitter (Derczynski et al., 2017) and all 6 sources of text in Ontonotes v5 (Hovy et al., 2006)[2], resulting in 8 domains. For SA, we follow

[1] Yahoo! answers, Email, NewsGroups, Reviews and Weblogs

[2] Broadcast News (BN), Broadcast Conversation (BC), Magazine (MZ), Telephone Conversation (TC) and Web data (WB)

Guo et al. (2020), selecting the same 5 categories[3] for experiments (Liu et al., 2017).

**Divergence Measures:** We consider 12 divergence measures. For Cos, we follow the instance based calculation (Ruder et al., 2017). For MMD, Wasserstein, and CORAL, we randomly sample 1000 sentences and average the results over 3 runs. For MMD, we experiment with different kernels (*cf.* Appendix A) and use default values of $\sigma$ from the GeomLoss package (Feydy et al., 2019). For TVO, KL-div, JS-div, Renyi-div, based on word frequency distribution we filter out stop words and consider the top 10k frequent words across domains to build our vocabulary (Ruder et al., 2017; Gururangan et al., 2020). We use $\alpha$=*0.99* for Renyi as found effective by Plank and van Noord (2011). We do not choose CE as it is mainly used in MT and not found effective for text classification and structured prediction (Ruder et al., 2017).

**Model Architecture:** For POS and NER, we follow the original BERT model where a linear layer is added upon the base DistilBERT model, and a prediction is made for every token. If the token is split into multiple tokens because of Byte Pair Encoding, the label for the first token is predicted. For sentiment analysis and domain discriminators, we pool the representation from the last layer of DistilBERT and add a linear layer for prediction. Grid search and hyper-parameter are set as in Appendix B.

## 5 Results

### 5.1 Performance Drop Correlation: Do traditional measures over distributions remain relevant?

For POS, the PAD measure has the best correlation with performance drop (cf. Table 1). Information-theoretic measures over word frequency distributions such as JS-div and KL-div, TVO, which have been prevalent for data selection and drop in performance (cf. Table 2) are comparable to PAD. Plank et al. (2014) claim that the errors in POS are dictated by out of vocabulary. The high correlations of KL-div, JS-div, and TVO with drop in performance corroborate their claims. For NER, MMD-RQ provides the best correlation of 0.495. CORAL – a higher-order measure and JS-div are comparable. For SA, Renyi-div and other information-theoretic measures provide considerably better correlation compared to higher-order measures. Cos is a widely used measure across applications. However, it *does not* provide a significant correlation for either task. TVO is used for selecting pretraining data for NER (Dai et al., 2019) and as a measure to gauge the benefits of fine-tuning pre-trained LMs on domain-specific data (Gururangan et al., 2020). Although TVO does not capture the nuances of domain divergences, it has strong, reliable correlations for performance drops. PAD has been suggested for data selection in SA by Ruder and Plank (2017) and for predicting drop in performance by Elsahar and Gallé (2019). Our analysis confirms that PAD provides good correlations with drop in performance across POS, NER, and SA.

We find no single measure to be superior across all tasks. However, information-theoretic measures consistently provide good correlations. Presently, as contextual word representations dictate results in NLP, simple measures based on frequency distributions are strong baselines to indicate the drop in performance. Although higher-order measures do not always provide the best correlation, they are differentiable, thus suited for end-to-end training of domain-invariant representations.

### 5.2 Domain Separation: Are different datasets really different domains?

Now that we have measured correlations, we want to analyse whether the datasets are indeed domains. And if so, how can we quantify it, and what can we conclude about its effect on performance? Aharoni and Goldberg (2020) and Ma et al. (2019) show that BERT representations reveal the underlying domains, although a few text segments from a dataset aptly belong to a different domain. But the degree to which the samples belong to a different domain for POS, NER, and SA is still unclear. We employ the Silhouette Coefficient (Rousseeuw, 1987), to check whether the datasets have underlying domains. We assume that all the data points from a dataset are in its own domain. A positive score indicates that datasets can be considered as well-separated domains; a negative score indicates that most of the points within a dataset can be assigned to a nearby domain; 0 indicates overlapping domains. For calculations, we use a subset of domain divergence measures that are metrics (a requirement for Silhouette scores) and can be calculated between single instances of text. We sample 200 points for each dataset as

[3] Apparel, Baby, Books, Camera and MR

| Measure | Correlations | | | Silhouette Coefficients | | |
|---|---|---|---|---|---|---|
| - | **POS** | **NER** | **SA** | **POS** | **NER** | **SA** |
| Cos | 0.018 | 0.223 | -0.012 | $-1.78 \times 10^{-1}$ | $-2.49 \times 10^{-1}$ | $-2.01 \times 10^{-1}$ |
| KL-Div | 0.394 | 0.384 | 0.715 | - | - | - |
| JS-Div | 0.407 | 0.484 | 0.709 | $-8.50 \times 10^{-2}$ | $-6.40 \times 10^{-2}$ | $+2.04 \times 10^{-2}$ |
| Renyi-Div | 0.392 | 0.382 | **0.716** | - | - | - |
| PAD | **0.477** | 0.426 | 0.538 | - | - | - |
| Wasserstein | 0.378 | 0.463 | 0.448 | $-2.11 \times 10^{-1}$ | $-2.36 \times 10^{-1}$ | $-1.70 \times 10^{-1}$ |
| MMD-RQ | 0.248 | **0.495** | 0.614 | $-4.11 \times 10^{-2}$ | $-3.04 \times 10^{-2}$ | $-1.70 \times 10^{-2}$ |
| MMD-Gaussian | 0.402 | 0.221 | 0.543 | $+4.25 \times 10^{-5}$ | $+2.37 \times 10^{-3}$ | $-8.42 \times 10^{-5}$ |
| MMD-Energy | 0.244 | 0.447 | 0.521 | $-9.84 \times 10^{-2}$ | $-1.14 \times 10^{-1}$ | $-8.48 \times 10^{-2}$ |
| MMD-Laplacian | 0.389 | 0.273 | 0.623 | $-1.67 \times 10^{-3}$ | $+4.26 \times 10^{-4}$ | $-1.08 \times 10^{-3}$ |
| CORAL | 0.349 | 0.484 | 0.267 | $-2.34 \times 10^{-1}$ | $-2.78 \times 10^{-1}$ | $-1.41 \times 10^{-1}$ |
| TVO | -0.437 | -0.457 | -0.568 | - | - | - |

Table 1: (l): Correlation of performance drops with divergence measures. Measures with higher correlations are better indicators of performance-drops. (r): Silhouette coefficients considering different domain divergence measures. The colours are from the taxonomy of divergence measures in Figure 1.

the time complexity increases exponentially with number of points. We average the results over 5 runs. We further obtained the t-SNE plots (Maaten and Hinton, 2008) for different divergence measures with DistilBERT representations (cf. Figures 3, 4 and 5 of Appendix D)

Almost all the measures across different tasks have negative values close to 0 (cf. Table 1, (r)). For POS – CORAL, Wasserstein, and Cos strongly indicate that text within a dataset belongs to other domains. However, for MMD-Gaussian, as seen in Fig 2a, we do observe some clustering, but the domains are overlapping. For NER, MMD-Gaussian and MMD-Laplacian indicate that the clusters overlap while all other metrics have negative values. For sentiment analysis, JS-div has positive values compared to other measures, and as seen in Figure 2c, we can see a better notion of distinct clusters. The Silhouette scores in tandem with the t-SNE plots indicate that datasets are, in fact, not distinct domains, and considering data-driven methods for defining domains is needed (Aharoni and Goldberg, 2020).

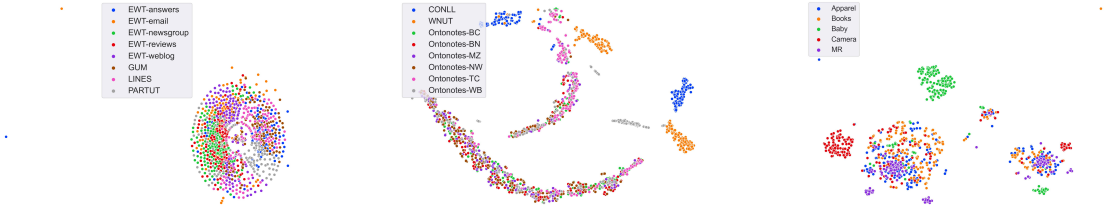
### 5.3 Discussion

One of the premises for drop in performance is that different datasets are from different domains. In Section 5.2, we showed that one has to treat the notion of *dataset-is-domain* carefully. Although we observe a drop in performance across different NLP tasks, it brings to light the question about the relationship between underlying domains and performance. We see better notions of clusters for NER and sentiment analysis (cf. Figures 2b and 2c). We can expect the drop in performance to be indicative of these domain separations. Comparing the best correlations from Table 1, we see higher correlations for NER and sentiment analysis, compared against POS. For POS, there are no indicative domain clusters and the effect of domain divergence on drop in performance may be less; whereas for SA, both the t-SNE plot and the Silhouette scores for JS-Div (cf. Figure 2c) corroborate comparatively better separation and we can see that the correlation is higher as well. To evaluate whether the domain adaptation techniques are bridging the differences in data distributions and the performance improvements are not because of model artifacts, one must be more careful in selecting the datasets.

Overlapping datasets also have consequences for data selection strategies. For example, Moore and Lewis (2010) select *pseudo in-domain data* from the source corpora. The Silhouette coefficients being negative and close to 0 shows that on average, many data points in a dataset belong to nearby domains. Data selection strategies thus may be effective. If the Silhouette coefficients are more negative and if more points in the source aptly belong to the target domain, we should expect increased sampling from such source domains to yield additional performance benefits in the target domain.

## Conclusion

We survey domain adaptation work, focusing on domain divergence measures and their usage for *data selection*, *learning domain-invariant representations*, and *making decisions in the wild*. We synthesised the divergence measures into a taxonomy of *information theoretic*, *geometric* and *higher-order* measures. While traditional measures are common for data selection and making




(a) (POS)-MMD-Gaussian (b) (NER)-Cos (c) (SA)-JS-Div

Figure 2: t-SNE plots for select measures. The complete set of diagrams are available in Appendix D.

decisions in the wild, higher-order measures are prevalent in learning representations. Based on our correlation experiments, silhouette scores, and t-SNE plots, we make the following recommendations:

- PAD is a reliable indicator of performance drop. Use it when there are sufficient examples to train a domain discriminator.
- JS-Div is symmetric and a formal metric. It is related to PAD, easy to compute, and serves as a strong baseline.
- While Cosine is popular, it is an unreliable indicator of performance drop.
- Do not consider datasets as domains for domain adaptation experiments. Instead, cluster the representations and define appropriate domains.

## References


Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. *ArXiv*, abs/2004.02105.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia. PMLR.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Eyal Ben-David, Carmel Rabinovitz, and Roi Reichart. 2020. Perl: Pivot-based domain adaptation for pretrained deep contextualized embedding models.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175.

Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 745–754, Honolulu, Hawaii. Association for Computational Linguistics.

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.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 343–351. Curran Associates, Inc.

Denny Britz, Quoc Le, and Reid Pryzant. 2017. Effective domain mixing for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 118–126, Copenhagen, Denmark. Association for Computational Linguistics.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2018. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 182–192, Brussels, Belgium. Association for Computational Linguistics.

Marine Carpuat, Yogarshi Vyas, and Xing Niu. 2017. Detecting cross-lingual semantic divergence for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*,

pages 69–79, Vancouver. Association for Computational Linguistics.

Boxing Chen, Colin Cherry, George Foster, and Samuel Larkin. 2017. Cost weighting for neural machine translation domain adaptation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 40–46, Vancouver. Association for Computational Linguistics.

Boxing Chen and Fei Huang. 2016. Semi-supervised convolutional networks for translation adaptation with tiny amount of in-domain data. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 314–323, Berlin, Germany. Association for Computational Linguistics.

Xilun Chen and Claire Cardie. 2018. Multinomial adversarial networks for multi-domain text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1226–1240, New Orleans, Louisiana. Association for Computational Linguistics.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ching-Yao Chuang, Antonio Torralba, and Stefanie Jegelka. 2020. Estimating generalization under distribution shifts via domain-invariant representations. *CoRR*, abs/2007.03511.

I. Csiszár. 1972. A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica*, 2(1):191–213.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2019. Using similarity measures to select pretraining data for NER. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1460–1470, Minneapolis, Minnesota. Association for Computational Linguistics.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. 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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

X. Ding, Q. Shi, B. Cai, T. Liu, Y. Zhao, and Q. Ye. 2019. Learning multi-domain adversarial neural networks for text classification. *IEEE Access*, 7:40323–40332.

Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. 2013. Adaptation data selection using neural language models: Experiments in machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 678–683, Sofia, Bulgaria. Association for Computational Linguistics.

Hady Elsahar and Matthias Gallé. 2019. To annotate or not? predicting performance drop under domain shift. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2163–2173, Hong Kong, China. Association for Computational Linguistics.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. 2019. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1180–1189. JMLR.org.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. 2015. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17:59:1–59:35.

Deepanway Ghosal, Devamanyu Hazarika, Abhinaba Roy, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2020. Kingdom: Knowledge-guided

domain adaptation for sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3198–3210. Association for Computational Linguistics.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J. Smola. 2007. A kernel method for the two-sample-problem. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press.

Shuhao Gu, Yang Feng, and Qun Liu. 2019. Improving domain adaptation translation with domain invariant and specific information. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3081–3091, Minneapolis, Minnesota. Association for Computational Linguistics.

Tao Gui, Qi Zhang, Haoran Huang, Minlong Peng, and Xuanjing Huang. 2017. Part-of-speech tagging for twitter with adversarial neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2411–2420, Copenhagen, Denmark. Association for Computational Linguistics.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2020. Multi-source domain adaptation for text classification via distancenet-bandits.

Jiang Guo, Darsh Shah, and Regina Barzilay. 2018. Multi-source domain adaptation with mixture of experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4694–4703, Brussels, Belgium. Association for Computational Linguistics.

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 for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, Hong Kong, China. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

James J. Jiang. 2007. A literature survey on domain adaptation of statistical classifiers.

Young-Bum Kim, Karl Stratos, and Dongchan Kim. 2017. Adversarial adaptation of synthetic or stale data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1297–1307, Vancouver, Canada. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. What's in a domain? learning domain-robust text representations using adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 474–479, New Orleans, Louisiana. Association for Computational Linguistics.

Le Liu, Yu Hong, Hao Liu, Xing Wang, and Jianmin Yao. 2014. Effective selection of translation model training data. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 569–573, Baltimore, Maryland. Association for Computational Linguistics.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–10, Vancouver, Canada. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Yajuan Lü, Jin Huang, and Qun Liu. 2007. Improving statistical machine translation performance by training data selection and optimization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 343–350, Prague, Czech Republic. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Anna Margolis. 2011. A literature review of domain adaptation with unlabeled data.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Minlong Peng, Qi Zhang, Yu-gang Jiang, and Xuanjing Huang. 2018. Cross-domain sentiment classification with target domain specific information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2505–2513, Melbourne, Australia. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Barbara Plank, Anders Johannsen, and Anders Søgaard. 2014. Importance weighting and unsupervised domain adaptation of POS taggers: a negative result. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 968–973, Doha, Qatar. Association for Computational Linguistics.

Barbara Plank and Gertjan van Noord. 2011. Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1566–1576, Portland, Oregon, USA. Association for Computational Linguistics.

Natalia Ponomareva and Mike Thelwall. 2012. Biographies or blenders: Which resource is best for cross-domain sentiment analysis? In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 488–499. Springer.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in nlp—a survey.

Sujith Ravi, Kevin Knight, and Radu Soricut. 2008. Automatic prediction of parser accuracy. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 887–896, Honolulu, Hawaii. Association for Computational Linguistics.

Robert Remus. 2012. Domain adaptation using domain similarity- and domain complexity-based instance selection for cross-domain sentiment analysis. In *12th IEEE International Conference on Data Mining Workshops, ICDM Workshops, Brussels, Belgium, December 10, 2012*, pages 717–723. IEEE Computer Society.

Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. 2020. A constructive prediction of the generalization error across scales. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Peter Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1):53–65.

Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2017. Data selection strategies for multi-domain sentiment analysis. *ArXiv*, abs/1702.02426.

Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with Bayesian optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 372–382, Copenhagen, Denmark. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Darsh Shah, Tao Lei, Alessandro Moschitti, Salvatore Romeo, and Preslav Nakov. 2018. Adversarial domain adaptation for duplicate question detection. In *Proceedings of the 2018 Conference on*

*Empirical Methods in Natural Language Processing*, pages 1056–1063, Brussels, Belgium. Association for Computational Linguistics.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. 2018. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4058–4065. AAAI Press.

Catarina Cruz Silva, Chao-Hong Liu, Alberto Poncelas, and Andy Way. 2018. Extracting in-domain training corpora for neural machine translation using data selection methods. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 224–231, Brussels, Belgium. Association for Computational Linguistics.

Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. QuEst - a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria. Association for Computational Linguistics.

J. Steinhardt. 2016. Unsupervised risk estimation with only structural assumptions.

Baochen Sun, Jiashi Feng, and Kate Saenko. 2017. Correlation alignment for unsupervised domain adaptation. In Gabriela Csurka, editor, *Domain Adaptation in Computer Vision Applications*, Advances in Computer Vision and Pattern Recognition, pages 153–171. Springer.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474.

Vincent Van Asch and Walter Daelemans. 2010. Using domain similarity for performance estimation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 31–36, Uppsala, Sweden. Association for Computational Linguistics.

Yogarshi Vyas, Xing Niu, and Marine Carpuat. 2018. Identifying semantic divergences in parallel text without annotations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1503–1515, New Orleans, Louisiana. Association for Computational Linguistics.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 560–566, Vancouver, Canada. Association for Computational Linguistics.

Yong Wang, Longyue Wang, Shuming Shi, Victor O.K. Li, and Zhaopeng Tu. 2019. Go from the general to the particular: Multi-domain translation with domain transformation networks. *ArXiv*, abs/1911.09912.

Zhenghui Wang, Yanru Qu, Liheng Chen, Jian Shen, Weinan Zhang, Shaodian Zhang, Yimei Gao, Gen Gu, Ken Chen, and Yong Yu. 2018. Label-aware double transfer learning for cross-specialty medical named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1–15, New Orleans, Louisiana. Association for Computational Linguistics.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. In *EMNLP*.

Guillaume Wisniewski and François Yvon. 2019. How Bad are PoS Tagger in Cross-Corpora Settings? Evaluating Annotation Divergence in the UD Project. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 218–227, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Dustin Wright and Isabelle Augenstein. 2020. Transformer based multi-source domain adaptation.

Yongfu Wu and Yuhong Guo. 2019. Dual adversarial co-learning for multi-domain text classification. *ArXiv*, abs/1909.08203.

Mengzhou Xia, Antonios Anastasopoulos, Ruochen Xu, Yiming Yang, and Graham Neubig. 2020. Predicting performance for natural language processing tasks.

Brian Xu, Mitra Mohtarami, and James R. Glass. 2019. Adversarial domain adaptation for stance detection. *ArXiv*, abs/1902.02401.

Keiji Yasuda, Ruiqiang Zhang, Hirofumi Yamamoto, and Eiichiro Sumita. 2008. Method of selecting training data to build a compact and efficient translation model. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Michihiro Yasunaga, Jungo Kasai, and Dragomir Radev. 2018. Robust multilingual part-of-speech tagging via adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 976–986, New Orleans, Louisiana. Association for Computational Linguistics.

Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. 2017. Central moment discrepancy (CMD) for domain-invariant representation learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Jiali Zeng, Jinsong Su, Huating Wen, Yang Liu, Jun Xie, Yongjing Yin, and Jianqiang Zhao. 2018. Multi-domain neural machine translation with word-level domain context discrimination. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 447–457, Brussels, Belgium. Association for Computational Linguistics.

Bing Zhao, Matthias Eck, and Stephan Vogel. 2004. Language model adaptation for statistical machine translation via structured query models. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 411–417, Geneva, Switzerland. COLING.

Joey Tianyi Zhou, Hao Zhang, Di Jin, Hongyuan Zhu, Meng Fang, Rick Siow Mong Goh, and Kenneth Kwok. 2019. Dual adversarial neural transfer for low-resource named entity recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3461–3471, Florence, Italy. Association for Computational Linguistics.

Suyang Zhu, Shoushan Li, and Guodong Zhou. 2019. Adversarial attention modeling for multi-dimensional emotion regression. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 471–480, Florence, Italy. Association for Computational Linguistics.

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, Vancouver, Canada. Association for Computational Linguistics.

Yftah Ziser and Roi Reichart. 2018. Pivot based language modeling for improved neural domain adaptation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1241–1251, New Orleans, Louisiana. Association for Computational Linguistics.

Yftah Ziser and Roi Reichart. 2019. Task refinement learning for improved accuracy and stability of unsupervised domain adaptation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5895–5906, Florence, Italy. Association for Computational Linguistics.

Bowei Zou, Zengzhuang Xu, Yu Hong, and Guodong Zhou. 2018. Adversarial feature adaptation for cross-lingual relation classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 437–448, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

## A Domain Divergence Measures

This section provides the necessary background on different kinds of divergence measures used in the literature. They can be either information-theoretic – which measure the distance between two probability distributions, geometric - which measure the distance between two vectors in a space, or higher-order which capture similarity in a projected space and consider higher order moments of random variables.

### A.1 Information-Theoretic Measures

Let $P$ and $Q$ be two probability distributions. These information-theoretic measures are used to capture differences between $P$ and $Q$.

**Kullback-Leibler Divergence (KL-Div)** $Q$ is called the reference probability distribution[4] more precisely KL is defined if only for all $Q(x)$ st $Q(x) = 0$, $P(x)$ is also 0. Undefined if $\exists$ x, $Q(x) = 0$ and $P(x) > 0$.

$$D_{KL}(P||Q) = \sum_x P(x) log\left(\frac{P(x)}{Q(x)}\right) \quad (1)$$

**Renyi Divergence (Renyi-Div)** Renyi Divergence is a generalisation of the KL Divergence and is also called $\alpha$-power divergence

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} log\left(\sum_x \frac{P(x)^\alpha}{Q(x)^{\alpha-1}}\right) \quad (2)$$

[4]KL divergence is asymmetric and cannot be considered a metric

| Paper | Task(s) | Information-Theoretic | | | | | Geometric | | Higher-Order | | | Others |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | KL | JS | Renyi | CE | Wass. | Cos | P-Norm | PAD | CMD | MMD | - |
| **DATA SELECTION** | | | | | | | | | | | | |
| (Plank and van Noord, 2011) | Par, POS | ✔ | ✔ | ✔ | | | ✔ | | | | | |
| (Dai et al., 2019) | NER | | | | | | | | | | | ✔ |
| (Ruder and Plank, 2017) | SA, NER, Par | | ✔ | ✔ | | | ✔ | ✔ | ✔ | | | ✔ |
| (Ruder et al., 2017) | SA | | ✔ | | | | ✔ | | ✔ | | | ✔ |
| (Remus, 2012) | SA | | ✔ | | | | | | | | | |
| (Lü et al., 2007) | SMT | | | | | | ✔ | | | | | |
| (Zhao et al., 2004) | SMT | | | | | | ✔ | | | | | |
| (Yasuda et al., 2008) | SMT | | | | ✔ | | | | | | | |
| (Moore and Lewis, 2010) | SMT | | | | ✔ | | | | | | | |
| (Axelrod et al., 2011) | SMT | | | | ✔ | | | | | | | |
| (Duh et al., 2013) | SMT | | | | ✔ | | | | | | | |
| (Liu et al., 2014) | SMT | | | | ✔ | | | | | | | |
| (van der Wees et al., 2017) | NMT | | | | ✔ | | | | | | | |
| (Silva et al., 2018) | NMT | | | | | | | | | | | ✔ |
| (Aharoni and Goldberg, 2020) | NMT | | | | | | ✔ | | | | | |
| (Wang et al., 2017) | NMT | | | | | | | ✔ | | | | |
| (Carpuat et al., 2017) | NMT | | | | | | | | | | | ✔ |
| (Vyas et al., 2018) | NMT | | | | | | | | | | | ✔ |
| (Chen and Huang, 2016) | SMT | | | | | | | | | | | ✔ |
| (Chen et al., 2017) | NMT | | | | | | | | | | | ✔ |
| **LEARNING REPRESENTATIONS** | | | | | | | | | | | | |
| (Ganin et al., 2015) | SA | | | | | | | | ✔ | | | |
| (Kim et al., 2017) | Intent-clf | | | | | | | | ✔ | | | |
| (Liu et al., 2017) | SA | | | | | | | | ✔ | | | |
| (Li et al., 2018) | Lang-ID | | | | | | | | ✔ | | | |
| (Chen and Cardie, 2018) | SA | | | | | | | | | ✔ | | |
| (Zellinger et al., 2017) | SA | | | | | | | | | ✔ | | |
| (Peng et al., 2018) | SA | | | | | | | | | ✔ | | |
| (Wu and Guo, 2019) | SA | | | | | | | | ✔ | | | |
| (Ding et al., 2019) | Intent-Clf | | | | | | | | ✔ | | | |
| (Shah et al., 2018) | Question sim | | | | | ✔ | | | ✔ | | | |
| (Zhu et al., 2019) | Emo-Regress | | | | | ✔ | | | | | | |
| (Gui et al., 2017) | POS | | | | | | | | ✔ | | | |
| (Zhou et al., 2019) | NER | | | | | | | | ✔ | | | |
| (Cao et al., 2018) | NER | | | | | | | | ✔ | | | |
| (Wang et al., 2018) | NER | | | | | | | | | | ✔ | |
| (Gu et al., 2019) | NMT | | | | | | | | ✔ | | | |
| (Britz et al., 2017) | NMT | | | | | | | | ✔ | | | |
| (Zeng et al., 2018) | NMT | | | | | | | | ✔ | | | |
| (Wang et al., 2019) | NMT | | | | | | | | ✔ | | | |
| **DECISIONS IN THE WILD** | | | | | | | | | | | | |
| (Ravi et al., 2008) | Parsing | | | | ✔ | | | | | | | |
| (Elsahar and Gallé, 2019) | SA, POS | | | | | | | | ✔ | | | ✔ |
| (Ponomareva and Thelwall, 2012) | SA | ✔ | ✔ | | | | ✔ | | | | | ✔ |
| (Van Asch and Daelemans, 2010) | POS | ✔ | | ✔ | | | | ✔ | | | | ✔ |

Table 2: Prior works using divergence measures for *Data Selection*, *Learning Representations* and *Decisions in the Wild*. Tasks can be *Par*: dependency parsing, *POS*: Parts of Speech tagging, *NER*: Named Entity Recognition, *SA*: Sentiment Analysis, *SMT*: Statistical and *NMT*: Neural Machine Translation, *Intent-Clf*: Intent classification, *Lang-ID*: Language identification, *Emo-Regress*: Emotional regression. *Wass.* denotes Wasserstein.

Here $\alpha \geq 0$ and $\alpha \neq 1$. Renyi divergence is equivalent to KL divergence as the limit $\alpha \to 1$.

**Jensen Shannon Divergence (JS-Div)** Jensen Shannon divergence (JS-divergence) is a symmetric version of KL-Divergence. It has many advantages. The square root of the Jensen Shannon Divergence is a metric and it can be used for non-continuous probabilities

$$D_{JS}(P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M)$$
$$M = \frac{1}{2}(P+Q) \tag{3}$$

**Entropy-Related - (CE)** Let, $H_T$, $H_S$ assigns entropy to a sentence using a language model trained on the target and source domain respectively. If $s$ is a text segment from the source domain, then the difference in entropy, as shown below gives the similarity of a source domain segment to the target domain. Some works just use $H_T$ and ignore $H_S$. Machine translation related works (Moore and Lewis, 2010), consider only the source language. Axelrod et al. (2011) extend to consider both languages – $src{-}lang$ and $trg{-}lang$ denote the two languages considered for machine translation, and this approach performs better for data selection. We present these variations in the formulae below and attribute the same name **CE** to both these variations in the literature review.

$$D_{CE} = H_T(s) - H_S(s) \tag{4}$$

$$D_{CE} = [H_T^{src-lang}(s) - H_S^{src-lang}(s)] + [H_T^{trg-lang}(s) - H_S^{trg-lang}(s)] \tag{5}$$

### A.2 Geometric Measures

Let $\vec{p}$ and $\vec{q}$ be two vectors in $\mathbb{R}^n$. Domain adaptation works use geometric metrics for continuous representations like word vectors.

**Cosine Similarity (Cos)**: It calculates the cosine of the angle between vectors. To measure the cosine distance between two points, we use $1 - Cos$

$$cos(\vec{p}, \vec{q}) = \frac{\vec{p}.\vec{q}}{\|p\| \, . \, \|q\|} \tag{6}$$

**$l_p$-norm (Norm)**:

**Euclidean distance or $l_2$ distance** measures the straight line distance between vectors and **Manhattan or $l_1$** measures the sum of the difference between their projections.

$$d_2(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{7}$$

$$d_1(p,q) = \sum_{i=1}^{n}|p_i - q_i| \tag{8}$$

### A.3 Higher-Order Measures

**$\mathcal{H}$-divergence and Proxy-A-Distance (PAD):** Ben-David et al. (2010) state that the error of a machine learning classifier in a target domain is bound by its performance on the source domain and the $\mathcal{H}$-divergence between the source and the target distributions. $\mathcal{H}$-divergence is expensive to calculate. An approximation of $\mathcal{H}$ is called *Proxy-A-Distance*. This definition has been adopted from (Elsahar and Gallé, 2019). Here $G : \mathcal{X} \to [0, 1]$ is a supervised machine learning model that classifies examples to the source and target domains, $D_s, D_t$. $|D|$ is the size of the training data and $\mathbb{1}$ is an indicator function

$$PAD = 1 - 2\epsilon(G_d) \tag{9}$$

$$\epsilon(G_d) = 1 - \frac{1}{|D|}\sum_{x_i \in D_s, D_t}|G(x_i) - \mathbb{1}(x_i \in D_s)| \tag{10}$$

**Wasserstein Distance:** Wasserstein Distance ( also called the Earth Mover's distance ) is another metric for two probability distributions. Intuitively, it measures the least amount of work done to transport probability mass from one probability distribution to another to make them equal. The work done in this case is measured as the mass transported multiplied by the distance of travel. It is known to be better than Kullback-Leibler Divergence and Jensen-Shannon Divergence when the random variables are high dimensional or otherwise. The Wasserstein metric is defined as

$$D_{Wasserstein} = \inf_{\gamma \in \pi}\sum_{x,y}\|x - y\|\,\gamma(x,y)$$

Here $\gamma \in \pi(P,Q)$ where $\pi(P,Q)$ is the set of all distributions where the marginals are $P$ and $Q$.

**Maximum Mean Discrepancy (MMD)**: MMD is a non-parametric method to estimate the distance between distributions based on Reproducing

Kernel Hilbert Spaces (RKHS). Given two random variables $X = \{x_1, x_2, ..., x_m\}$ and $Y = \{y_1, y_2, ...., y_n\}$ that are drawn from distributions $P$ and $Q$, the empirical estimate of the distance between distribution $P$ and $Q$ is given by

$$MMD(X,Y) = \left\| \left(\frac{1}{m}\sum_{i=1}^{m}\phi(x_i) - \frac{1}{n}\sum_{i=1}^{n}\phi(y_i)\right)\right\|_{\mathcal{H}} \tag{11}$$

Here $\phi : \mathcal{X} \rightarrow \mathcal{H}$ are nonlinear mappings of the samples to a feature representation in a RKHS. In this work, we map the contextual word representations of the text to RKHS. The different kinds of kernels we use in this work are given below. We use the default values of $\sigma = 0.05$ of the GeomLoss package (Feydy et al., 2019).

**Rational Quadratic Kernel**

$$\phi(x,y) = \left(1 + \frac{1}{2\alpha}(x-y)^T\Theta^{-2}(x-y)\right)^{-\alpha}$$

**Energy**

$$\phi(x,y) = -\left\|x-y\right\|_2$$

**Gaussian**

$$\phi(x,y) = exp(-\frac{\left\|x-y\right\|_2^2}{2\sigma^2})$$

**Laplacian**

$$\phi(x,y) = exp(-\frac{\left\|x-y\right\|_2}{\sigma})$$

**Correlation Alignment (CORAL)**: Correlation alignment is the distance between the second-order moment of the source and target samples. If $d$ is the representation dimension, $\|\|_F$ represents Frobenius norm and $Cov_S, Cov_T$ is the covariance matrix of the source and target samples, then CORAL is defined as

$$D_{CORAL} = \frac{1}{4d^2}\left\|Cov_S - Cov_T\right\|_F^2 \tag{12}$$

**Central Moment Discrepancy (CMD)**: Central Moment Discrepancy is another metric that measures the distance between source and target distributions. It not only considers the first moment and second moment, but also other higher-order moments. While MMD operates in a projected space, CMD operates in the representation space. If $P$ and $Q$ are two probability distributions and $X = \{X_1, X_2, ...., X_N\}$ and $Y = \{Y_1, Y_2, ...., Y_N\}$ are random vectors that are independent and identically distributed from $P$ and $Q$ and every component of the vector is bounded by $[a, b]$, CMD is then defined by

$$CMD(P,Q) = \frac{1}{|b-a|}\left\|E(X) - E(Y)\right\|_2 + \sum_{k=2}^{\infty}\frac{1}{|b-a|^k}\left\|c_k(X) - c_k(Y)\right\|_2 \tag{13}$$

where $E(X)$ is the expectation of X and $c_k$ is the $k-th$ order central moment which is defined as

$$c_k(X) = E\left(\prod_{i=1}^{N}(X_i - E(X_i))^{r_i}\right) \tag{14}$$

and $r_1 + r_2 + r_N = k$ and $r_1....r_N \geq 0$

### A.4 Other Measures

**Bhattacharya Coefficient:** If $P$ and $Q$ are probability distributions, then the Bhattacharya coefficient and Bhattacharya distance are defined as

$$Bhattacharya(P,Q) = \sum_x \sqrt{P(x)Q(x)} \tag{15}$$

$$D_{Bhattacharya} = -log(Bhattacharya(P,Q)) \tag{16}$$

**Term Vocabulary Overlap (TVO)**: This measures the proportion of target vocabulary that is also present in the source vocabulary. If $V_S$ is the source domain vocabulary and $V_T$ is the target domain vocabulary, then the Term Vocabulary Overlap between the source domain ($D_S$) and the target domain ($D_T$) is given by

$$TVO(D_S, D_T) = \frac{|V_S \bigcap V_T|}{|V_T|} \tag{17}$$

**Word Vector Variance**: Different contexts in which a word is used in two different datasets can be used as an indication of the divergence between two datasets. Let $\vec{w}^i_{src}$ denote the word embedding of word $i$ in source domain and $\vec{w}^i_{trg}$ is the word embedding of the same word in the target domain. Let $d$ be the dimension of the word embedding.

The word vector variance between the source domain ($D_S$) and the target domain ($D_T$) is given by

$$WVV(D_S, D_T) = \frac{1}{|V_S| * d} \sum_{i}^{|V_s|} \left\| w^i_{src} - w^i_{trg} \right\|_2^2 \tag{18}$$

## B Model Hyperparameters

For POS, NER and Sentiment Analysis models, we do a grid search of learning rate in {*1e-01, 1e-05, 5e-05*} and dropout in $\{0.2, 0.3, 0.4, 0.5\}$ and number of epochs in $\{25, 50\}$. PAD requires a domain discriminator. We sample as many samples in the target domain as the source domain (Ruder et al., 2017) and train a DistilBERT based classifier. For every domain discriminator we do a grid search of learning rate in {*1e-05, 5e-05*}, dropout in $\{0.4, 0.5\}$ and number of epochs in $\{10, 25\}$. For POS and NER we monitor the macro F-Score and for domain discrimination we monitor the accuracy scores and chose the best model after the grid search for all subsequent calculations. For training the models we use the Adam Optimiser (Kingma and Ba, 2014) with the $\beta_1 = 0.9$ and $\beta_2 = 0.99$ and $\epsilon$ as 1e-8. We use HuggingFace Transformers (Wolf et al., 2019) for all our experiments.

## C Cross-Domain Performances

### C.1 Parts of speech tagging

Table 3 shows the hyper parameters for the best model for POS and Table 4 shows the cross domain performances.

### C.2 Named Entity Recognition

Table 5 shows the hyper parameters for the best model for NER and Table 6 shows the cross domain performances.

### C.3 Sentiment Analysis

Table 7 shows the hyper parameters for the best model for Sentiment analysis and Table 8 shows the cross domain performances.

## D t-SNE Plots

We plot the t-SNE plots for POS (Figure 3), NER (Figure 4) and Sentiment Analysis (SA) (Figure 5). We sample 200 points from each of the datasets for the plot.

| Dataset | Epochs | Learning Rate | Dropout | Fscore |
|---|---|---|---|---|
| EWT-answers | 50 | $5 \times 10^{-5}$ | 0.4 | 95.38 |
| EWT-email | 50 | $1 \times 10^{-5}$ | 0.3 | 96.62 |
| EWT-newsgroup | 50 | $5 \times 10^{-5}$ | 0.5 | 95.92 |
| EWT-reviews | 50 | $5 \times 10^{-5}$ | 0.4 | 96.97 |
| EWT-weblog | 50 | $5 \times 10^{-5}$ | 0.3 | 97.03 |
| GUM | 50 | $1 \times 10^{-5}$ | 0.3 | 95.73 |
| LINES | 50 | $5 \times 10^{-5}$ | 0.3 | 97.38 |
| PARTUT | 50 | $1 \times 10^{-5}$ | 0.4 | 97.06 |

Table 3: Model performance and hyper-parameters producing the best model for Parts of Speech Tagging trained using DistilBERT as the base model. The datasets are from the Universal Dependencies Corpus (UD) (Nivre et al., 2016). 5 corpora are from the English Word Tree (EWT) portion which are EWT-answers, EWT-email, EWT-newgroup, EWT-reviews, EWT-weblog, GUM, LINES, PARTUT

| Source/Target | EWT-answers | EWT-email | EWT-newsgroup | EWT-reviews | EWT-weblog | GUM | LINES | PARTUT |
|---|---|---|---|---|---|---|---|---|
| EWT-answers | 95.38 | 93.96 | 94.02 | 95.83 | 95.64 | 93.58 | 93.86 | 92.06 |
| EWT-email | 94.11 | 96.62 | 94.40 | 95.42 | 95.37 | 93.08 | 93.98 | 93.47 |
| EWT-newsgroup | 94.71 | 95.07 | 95.92 | 95.31 | 96.80 | 93.82 | 93.83 | 92.74 |
| EWT-reviews | 94.99 | 94.51 | 94.56 | 96.97 | 95.55 | 93.07 | 94.27 | 92.62 |
| EWT-weblog | 95.38 | 93.96 | 94.02 | 95.83 | 95.64 | 93.58 | 93.87 | 92.06 |
| GUM | 91.63 | 92.59 | 91.75 | 93.55 | 93.56 | 95.73 | 93.54 | 93.12 |
| LINES | 89.79 | 89.77 | 88.76 | 92.39 | 90.77 | 91.75 | 97.38 | 92.68 |
| PARTUT | 89.27 | 89.54 | 89.56 | 91.28 | 92.27 | 90.65 | 92.97 | 96.65 |

Table 4: Cross-domain performance for POS tagging. The best model for each source domain is tested on the test dataset of the same domain and all other domains.

| Dataset | Epochs | Learning Rate | Dropout | Fscore |
|---|---|---|---|---|
| CONLL-2003 | 50 | $5 \times 10^{-5}$ | 0.5 | 0.90 |
| WNUT | 25 | $5 \times 10^{-5}$ | 0.5 | 0.50 |
| Onto-BC | 50 | $5 \times 10^{-5}$ | 0.5 | 0.82 |
| Onto-BN | 50 | $1 \times 10^{-5}$ | 0.3 | 0.89 |
| Onto-MZ | 50 | $1 \times 10^{-5}$ | 0.3 | 0.86 |
| Onto-NW | 25 | $5 \times 10^{-5}$ | 0.4 | 0.89 |
| Onto-TC | 50 | $1 \times 10^{-5}$ | 0.5 | 0.75 |
| Onto-WB | 50 | $5 \times 10^{-5}$ | 0.4 | 0.63 |

Table 5: Model performance for Named Entity Recognition trained using DistilBERT as the base model. The datasets are CONLL-2003, Emerging and Rare Entity Recognition twitter dataset (WNUT), and six different sources of text in Ontonotes v5 (Hovy et al., 2006)

| Source/Target | CONLL-2003 | WNUT | ONTO-BC | ONTO-BN | ONTO-MZ | ONTO-NW | ONTO-TC | ONTO-WB |
|---|---|---|---|---|---|---|---|---|
| CONLL-2003 | 0.90 | 0.37 | 0.54 | 0.65 | 0.59 | 0.54 | 0.51 | 0.41 |
| WNUT | 0.66 | 0.50 | 0.40 | 0.44 | 0.49 | 0.42 | 0.49 | 0.33 |
| ONTO-BC | 0.48 | 0.31 | 0.82 | 0.81 | 0.77 | 0.74 | 0.72 | 0.45 |
| ONTO-BN | 0.53 | 0.37 | 0.77 | 0.89 | 0.76 | 0.79 | 0.76 | 0.47 |
| ONTO-MZ | 0.49 | 0.29 | 0.72 | 0.78 | 0.86 | 0.75 | 0.69 | 0.45 |
| ONTO-NW | 0.52 | 0.32 | 0.73 | 0.86 | 0.73 | 0.89 | 0.76 | 0.46 |
| ONTO-TC | 0.51 | 0.37 | 0.61 | 0.64 | 0.57 | 0.55 | 0.75 | 0.41 |
| ONTO-WB | 0.43 | 0.12 | 0.52 | 0.63 | 0.54 | 0.57 | 0.52 | 0.63 |

Table 6: Cross-domain performance for NER. The best model for each source domain is tested on the test dataset of the same domain and all other domains.

| Dataset | Epochs | Learning Rate | Dropout | Fscore |
|---|---|---|---|---|
| Apparel | 25 | $1 \times 10^{-5}$ | 0.4 | 91.25 |
| Baby | 50 | $5 \times 10^{-5}$ | 0.4 | 93.75 |
| Books | 50 | $1 \times 10^{-5}$ | 0.4 | 92 |
| Camera/Photo | 25 | $1 \times 10^{-5}$ | 0.4 | 92 |
| MR | 50 | $5 \times 10^{-5}$ | 0.3 | 82.5 |

Table 7: Model performance for Named Entity Recognition trained using DistilBERT as the base model. We chose 5 out of 16 datasets from (Liu et al., 2017) which are Apparel, Baby, Books, Camera/Photo, and MR.

| **Source/Target** | Apparel | Baby | Books | CameraPhoto | MR |
|---|---|---|---|---|---|
| Apparel | 0.91 | 0.9100 | 0.85 | 0.87 | 0.77 |
| Baby | 0.89 | 0.9375 | 0.86 | 0.89 | 0.75 |
| Books | 0.88 | 0.8875 | 0.92 | 0.87 | 0.79 |
| CameraPhoto | 0.89 | 0.89 | 0.86 | 0.92 | 0.75 |
| MR | 0.76 | 0.76 | 0.8375 | 0.74 | 0.83 |

Table 8: Cross-domain performance for Named Entity Recognition. The best model for each source domain is tested on the test dataset of the same domain and all other domains.

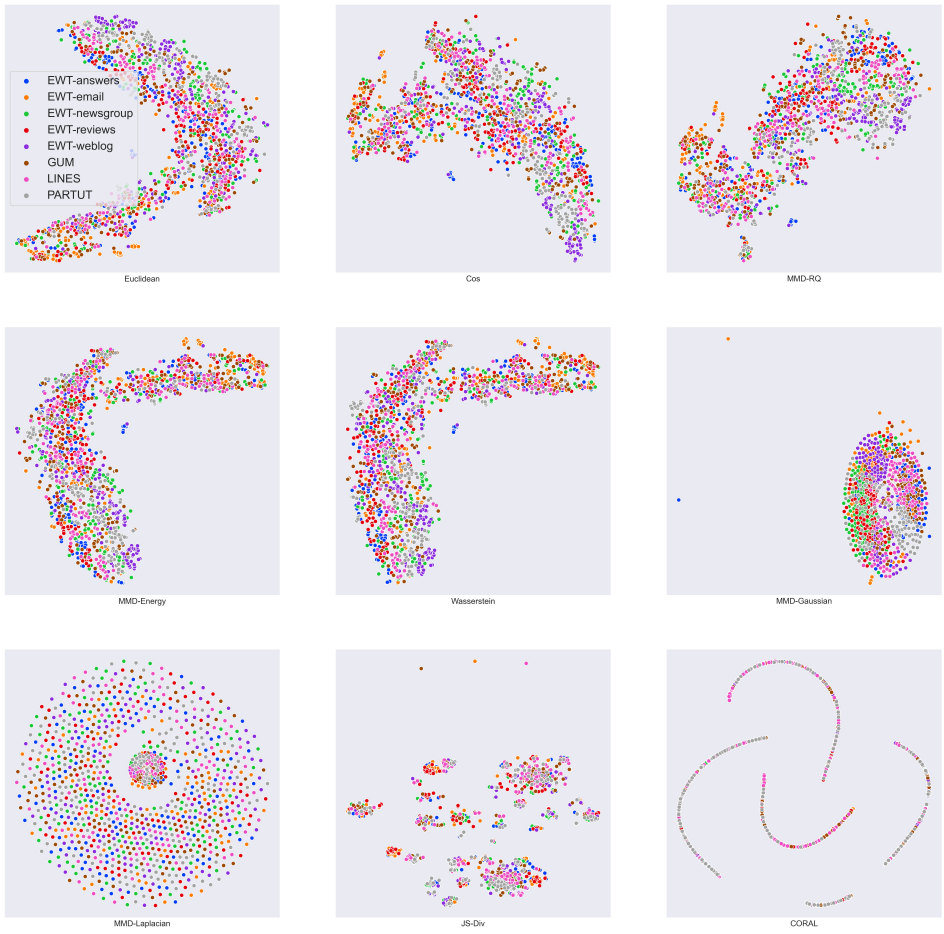



Figure 3: t-SNE plots for POS corpora.

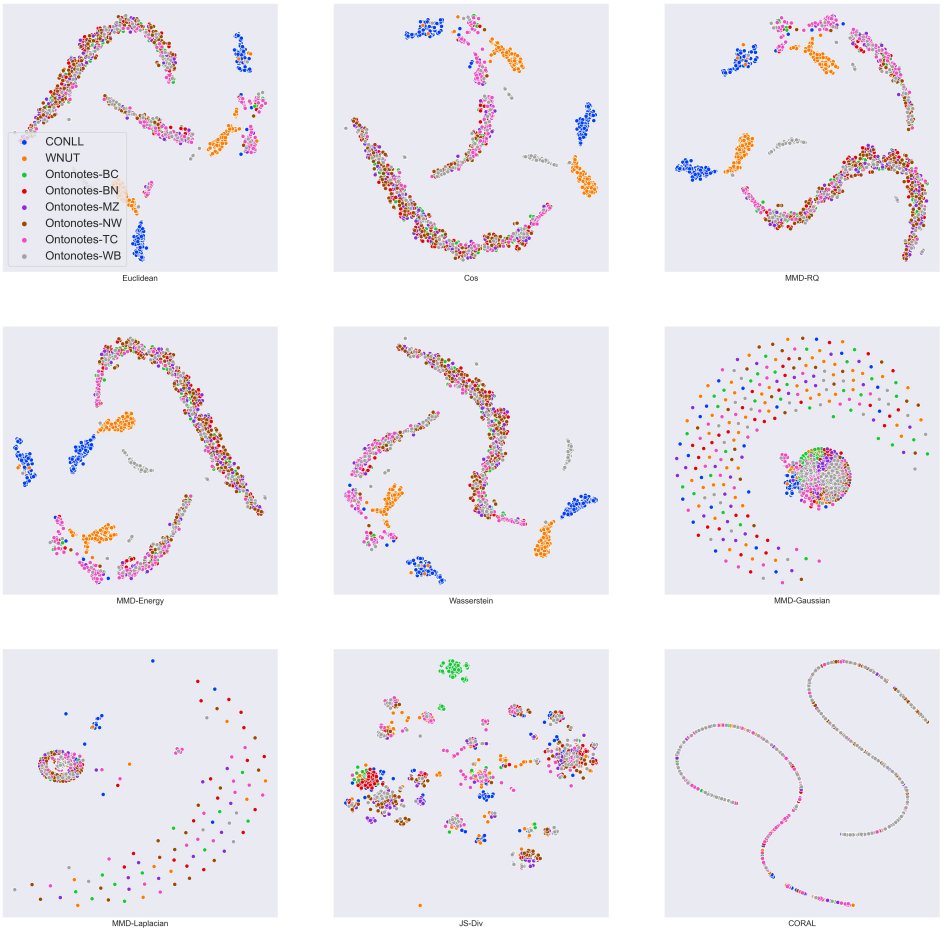



Figure 4: t-SNE plots for NER corpora.

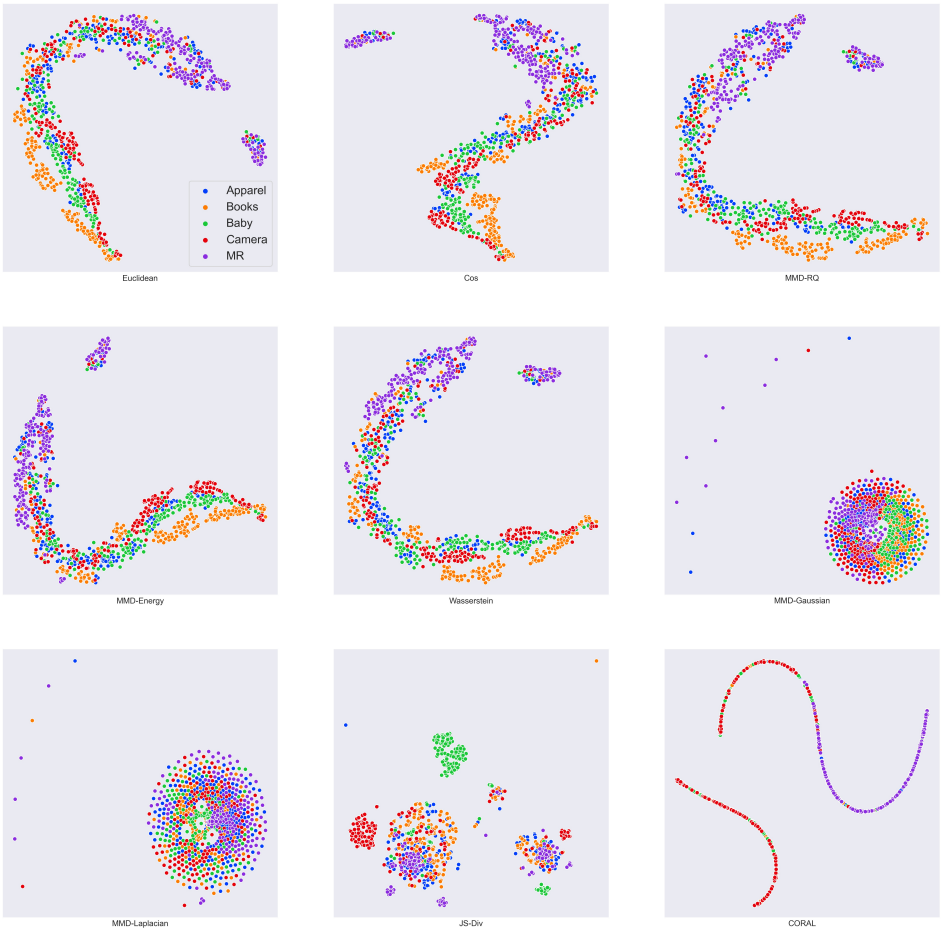



Figure 5: t-SNE plots for SA corpora.