Domain Adaptation for Arabic Cross-Domain and Cross-Dialect Sentiment Analysis from Contextualized Word Embedding

Abdellah El Mekki[1] **Abdelkader El Mahdaouy**[1]
Ismail Berrada[1] **Ahmed Khoumsi**[2]
[1]School of Computer Sciences, Mohammed VI Polytechnic University, Morocco
[2]Dept. Electrical & Computer Engineering, University of Sherbrooke, Canada
{firstname.lastname}@um6p.ma
ahmed.khoumsi@usherbrooke.ca

Abstract

Finetuning deep pre-trained language models has shown state-of-the-art performances on a wide range of Natural Language Processing (NLP) applications. Nevertheless, their generalization performance drops under domain shift. In the case of Arabic language, diglossia makes building and annotating corpora for each dialect and/or domain a more challenging task. Unsupervised Domain Adaptation tackles this issue by transferring the learned knowledge from labeled source domain data to unlabeled target domain data. In this paper, we propose a new unsupervised domain adaptation method for Arabic cross-domain and cross-dialect sentiment analysis from Contextualized Word Embedding. Several experiments are performed adopting the coarse-grained and the fine-grained taxonomies of Arabic dialects. The obtained results show that our method yields very promising results and outperforms several domain adaptation methods for most of the evaluated datasets. On average, our method increases the performance by an improvement rate of 20.8% over the zero-shot transfer learning from BERT.

1 Introduction

The Arabic language is characterized by two main language varieties: Modern Standard Arabic (MSA) and Arabic dialect. MSA has a standard written form and acquires an official status across the Arab countries, while Dialectal Arabic refers to the informal spoken dialects in the Arab World (Habash, 2010). These dialects are used in daily life but have no standard written form (Saadane and Habash, 2015; Habash et al., 2018; Eryani et al., 2020). Geographically and according to (Zaidan and Callison-Burch, 2014), Arabic dialects can be classified into five coarse-grained regional dialects: Egyptian, Levantine, Gulf, Iraqi, and Maghrebi. Recent studies have categorized dialectal Arabic into more fine-grained levels, including countries and cities (Bouamor et al., 2019; Muhammad et al., 2020). These dialects differ from one another and from MSA, to a varying degree, at different linguistic levels (Salameh et al., 2018; Erdmann et al., 2018).

With the unprecedented reach of social media platforms, Sentiment Analysis (SA) has become a fundamental task for many applications. Most research work in this area has been devoted to English and other European languages, while some research studies have addressed the question of transfer learning from MSA to dialectal Arabic. However, Khaddaj et al. (2019) and Qwaider et al. (2019) have shown that zero-shot transfer learning, from models trained on MSA data, does not perform well for SA on dialectal Arabic data. So, existing works have focused on building resources and annotating corpora for a few dialects where most of them were collected from social media (Medhaffar et al., 2017; Al-Twairesh et al., 2017; Baly et al., 2018; Moudjari et al., 2020; Oueslati et al., 2020). Nevertheless, dealing with Arabic dialects as standalone languages is challenging since building manually such resources is costly and time-consuming.

It is well known that the generalization performance of Machine Learning (ML) models drops in the case of domain shift (out of distribution data). Hence, there is an imperative need to leverage existing labeled data from other related domains, in order to address this challenge. The aim is to accurately transfer the learned knowledge from a source domain labeled data to a new target domain data. On the one hand, adaptive pretraining of contextualized word embedding models has shown an effective transfer learning performance under domain shift (Han and Eisenstein, 2019; Rietzler et al., 2020). It consists of finetuning the pre-trained language models on large unlabeled corpus from the target domain using the MLM objective. On the other hand, self-training and domain-adversarial learning have been applied

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2824–2837
June 6–11, 2021. ©2021 Association for Computational Linguistics

successfully to many NLP applications (Li et al., 2020; Ramponi and Plank, 2020; Ye et al., 2020; Ganin et al., 2016). An effective method that combines domain-adversarial training and self-training is the Adversarial-Learned Loss for Domain Adaptation (ALDA) (Chen et al., 2020). The domain-adversarial training aligns both domains' output distributions, while self-training captures the discriminative features of the target domain data.

In this paper, we introduce a new unsupervised domain adaptation method for Arabic cross-domain and cross-dialect sentiment analysis based on AraBERT language model (Antoun et al., 2020) and the Adversarial-Learned Loss for Domain Adaptation (ALDA) (Chen et al., 2020). Due to limited amount of unlabeled data for most target domains-dialects, we do not rely on the adaptive pre-training of AraBERT model. Our method leverages the potentials of: i) contextualized word embeddings to learn high-level text representation, ii) adversarial domain training to match the output distributions of domains and dialects, and iii) self-training to capture the discriminative features of the target domain data.
To summarize, our main contributions are as follows:

- The proposition of a new unsupervised domain adaptation method for Arabic SA.
- The study of three possible challenging scenarios of domain adaptation for Arabic SA.
- The achievement of very promising results on several Arabic cross-domain and cross-dialect sentiment classification datasets.

To the best of our knowledge, this is the first study that investigates domain adaptation for cross-domain, cross-dialect and cross-domain & cross-dialect sentiment analysis, adopting the coarse-grained and the fine-grained taxonomies of Arabic dialects. The proposed method outperforms several state-of-the-art methods on most test datasets.

The rest of this paper is organized as follows. Section 2 presents related work. In Section 3, we introduce our method. Section 4 illustrates the conducted experiments, and discusses the obtained results. Finally, in Section 6, we conclude the paper and outline a few directions for future work.

2 Related Work

Arabic sentiment classification. Recently, tangible progress has been made for Arabic sentiment analysis (Badaro et al., 2019; Al-Ayyoub et al., 2019). This has been achieved by publishing datasets (Elnagar et al., 2018; Ashraf and Omar, 2016; Aly and Atiya, 2013; ElSahar and El-Beltagy, 2015; Nabil et al., 2015), sentiment lexicons (Badaro et al., 2014; El-Beltagy, 2016; Gilbert Badaro and Habash, 2018), and proposing models as well as architectures that reach decent accuracy scores (Al Sallab et al., 2015; Antoun et al., 2020; Abdul-Mageed et al., 2020). As an example, the pre-trained language model AraBERT (Antoun et al., 2020) has achieved state-of-the-art performance on Arabic sentiment classification tasks. Nevertheless, most of these achievements are still limited to the MSA, and to some Arabic dialects and domains (Badaro et al., 2019; Al-Ayyoub et al., 2019).

Unsupervised domain adaptation. In the past few years, there has been considerable interest in unsupervised domain adaptation for cross-domain NLP tasks, including cross-domain sentiment analysis (Ramponi and Plank, 2020). Previous work has focused on minimizing the discrepancy between domains by aligning the output distributions of the source and the target domains. Maximum Mean Discrepancy (MMD) (Gretton et al., 2012), KL-divergence (Zhuang et al., 2015), Correlation Alignment (CORAL) (Sun and Saenko, 2016), and domain-adversarial learning (Ganin et al., 2016) are among the most widely used methods to learn domain-invariant features. In the same vein, other researchers have adopted self-training approach in order to learn discriminative features of the target domain (Ramponi and Plank, 2020; Ye et al., 2020). The latter approach enables the model to be also trained on some samples of the target domain. The main idea is to select a subset of pseudo-labels, predicted on the target domain inputs, for which the model's confidence is higher than a fixed threshold, and to incorporate them into the model loss. However, pseudo-labels are generally noisy and may hurt the performance of the model. Chen et al. (2020) have tackled this issue by introducing the adversarial-learned loss for domain adaptation where the discriminator corrects the noise in the pseudo-labels by generating noise vectors that are specific for each domain.

Domain adaptation for cross-domain sentiment analysis. In order to learn cross-domain text representation, several domain adaptation methods have relied on pivot features extraction. Inspired from structural correspondence learning, Yu and Jiang (2016) have proposed a method to learn continuous sentence embedding employing CNN model across various domains. Li et al. (2018) have introduced a domain adaption method which can be extended to documents. The latter method uses a hierarchical attention transfer network for extracting the pivots and non-pivots features between source and target domains. Ziser and Reichart (2018) have proposed language modeling objective to learn a model scratch rather than adapting a pre-trained embedding model.

Recently, several methods have been introduced for domain adaptation based on adaptive pre-training of contextualized word embeddings (Han and Eisenstein, 2019; Li et al., 2020; Vu et al., 2020). The latter approach relies on the availability of a large amount of unlabeled data in the target domain to finetune/adapt the existing pre-trained language model using the MLM objective. Rietzler et al. (2020) have proposed an unsupervised domain adaptation method for aspect-target sentiment classification based on BERT adaptive pre-training. Vu et al. (2020) have presented an adaptive pre-training method that adversarially masks out tokens that are hard to be reconstructed by the MLM. In another work, (Du et al., 2020) have proposed to combine BERT domain-aware training and adversarial-domain learning (Ganin et al., 2016) for cross-domain sentiment analysis. The domain-aware training combines the adaptive pre-training using the MLM objective and a Domain Distinguish Task (DDT). For cross-domain and cross-lingual domain adaptation, Li et al. (2020) have introduced an unsupervised feature decomposition method based on the mutual information to extract domain-invariant and domain-specific features using the XLM language model (Lample and Conneau, 2019).

For the Arabic language, Khaddaj et al. (2019) have introduced a domain adaptation method for cross-domain and cross-dialect sentiment analysis, combining domain adversarial training (Ganin et al., 2016) with denoising autoencoder for representation learning. The input sentences of both domains are represented using the bag-of-words representation by selecting the top 5,000 most frequent unigrams and bigrams. The obtained results on the Levantine multi-topic ArSentD-LEV dataset (Baly et al., 2018) show that combining the reconstruction loss with the adversarial training has slightly improved the performance in some cases. Nevertheless, the overall obtained results show that the zero-shot transfer from the SVM model achieves competitive results for some datasets. In another work, Qwaider et al. (2019) have shown that models that are trained on MSA for the task of sentiment classification generalize poorly to dialectal Arabic data. For improving the results, they have performed domain adaptation using feature engineering and sentiment lexicons.

3 Method

In this section, we present our model architecture. The noise-correcting discriminator, the classifier and the generator losses, employed in our model, are those of ALDA model (Chen et al., 2020).

3.1 Model architecture

In unsupervised domain adaptation settings, for sentiment analysis, we are given a labeled source domain $\mathcal{D}_S = \{(x_s^i, y_s^i)\}_{i=1}^{n_s}$ having K classes and an unlabeled target domain $\mathcal{D}_T = \{x_t^i\}_{i=1}^{n_t}$. The aim is then to transfer the learned knowledge from $\mathcal{D}_S$ to $\mathcal{D}_T$. In other words, the objective is to train a robust classifier using the labeled source domain data that generalizes well on the target domain test data.

Figure 1 presents the general framework of our method. We aim to leverage the strength of both the domain adversarial training and the self-training in a unified framework. The adversarial training aligns both domains' output distributions, whereas the self-training considers the discriminative features of the target domain. Besides, AraBERT is used as a generator to extract high-level representation from both source and target domains sentences.

The generator G, the AraBERT encoder, is trained to extract features from the input sentences for both domains: $h_{[CLS]} = G(x)$ corresponds to the hidden state of the $[CLS]$ token. The weights of the generator are shared between both domain inputs.

The classifier C operates on the hidden states $h_{[CLS]}$ to classify the input instances x and outputs a probability vector $p(y = k|x) = Softmax(W_c h_{[CLS]} + b_c)$ for both domains (p_s and p_t), where b_c and W_c are the bias vector and the weight matrix on the classification layer, respectively.

The generator G tries to confuse the discrimina-

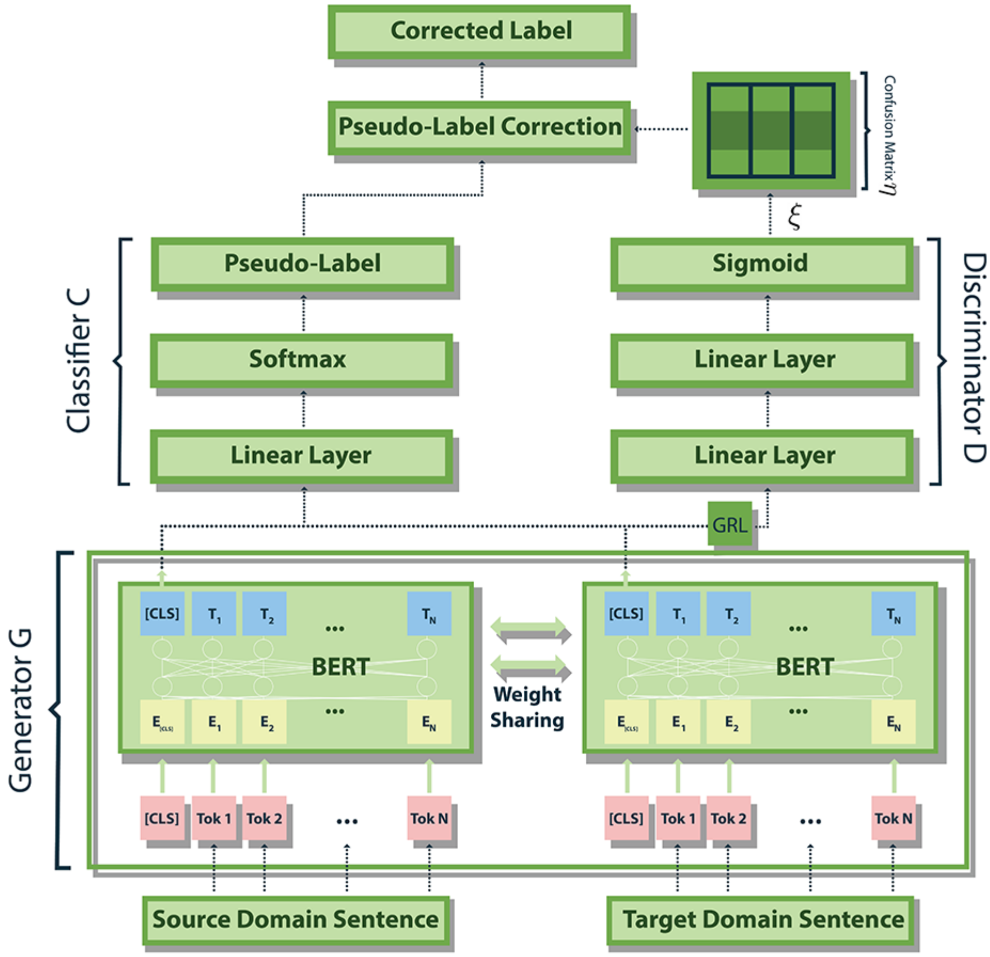

Figure 1: The general framework of our method

tor D by maximizing its loss. Thus, the generator aligns both domains' output distributions, whereas the discriminator must distinguish both domain features by generating different noise vectors for each domain. These noise vectors are employed to correct the pseudo-labels predicted by the classifier C. The Gradient Reversal Layer (GRL) reverses the gradient of the discriminator's loss during the back-propagation step.

3.2 Noise correcting discriminator

The input of the discriminator D corresponds to the hidden state $h_{[CLS]}$ of the generator G. D is trained to produce a noise vector $\xi^{(x)} = \sigma(D(h_{[CLS]}))$ by applying σ, the $Sigmoid$ activation, on its output layer. Note that, the output layer size is equal to K, the number of classes. Each component of the noise vector estimates the probability that the predicted label is the correct label $\xi_k^{(x)} = p(y = k|\hat{y} = k, x)$. Hence, instead of being trained to classify the source domain sentences and the ones of the target domain, G is trained to generate different noise vectors for each domain. The noise vector is used to estimate the confusion matrix $\eta = (\eta_{kl})$, which is applied to correct the target domain's pseudo-labels, predicted by the classifier C. The intuition behind the ALDA model is that, if we appropriately estimate the confusion matrix, the noise in the pseudo-labels predicted by the classifier, can be efficiently corrected (Chen et al., 2020).

Assuming that the noise in pseudo-labels is class-wise uniform with vector $\xi_k^{(x_t)}$, the confusion matrix is then given by:

$$\eta_{kl}^{x_t} = \begin{cases} \xi_k^{(x_t)} & \text{if } k = l \\ \frac{1-\xi_l^{(x_t)}}{K-1} & \text{else} \end{cases}$$

The corrected label vector in the target domain is given by $c^{(x_t)} = \sum_l \eta_{kl}^{x_t} p(\hat{y}_t = l|x_t)$ (l is the predicted pseudo label). For the source domain, the corrected label vector $c^{(x_s)}$ is computed using the same procedure.

For the source domain, the discriminator minimizes the binary cross-entropy loss $\mathcal{L}_{bce}$ between the corrected label vectors and the ground truth labels y_s:

$$\mathcal{L}_{adv}(x_s, y_s) = \mathcal{L}_{bce}(c^{(x_s)}, y_s) \tag{1}$$

For the target domain, the discriminator minimizes the binary cross-entropy loss $\mathcal{L}_{bce}$ between the corrected label vector and the opposite distribution of the predicted pseudo-label $u^{(\hat{y}_t)}$:

$$\mathcal{L}_{adv}(x_t) = \mathcal{L}_{bce}(c^{(x_t)}, u^{(\hat{y}_t)}) \tag{2}$$

where $u^{(\hat{y}_t)}$ is computed as follows:

$$u_k^{(\hat{y}_t)} = \begin{cases} 0 & \text{if } \hat{y}_t = k \\ \frac{1}{K-1} & \text{else} \end{cases}$$

To discriminate between both domains, the discriminator minimizes the following total adversarial loss:

$$\mathcal{L}_{adv}(x_s, y_s, x_t) = \mathcal{L}_{adv}(x_s, y_s) + \mathcal{L}_{adv}(x_t) \tag{3}$$

In order to make the training more stable, ALDA incorporates the classification loss of the source domain as a regularization term into the discriminator. Thus, the discriminator must also correctly classify the source domain data. The regularization term is given by:

$$\mathcal{L}_{reg}(x_s, y_s) = \mathcal{L}_{ce}(p_d^{(x_s)}, y_s) \tag{4}$$

where $p_d^{(x_s)} = Softmax(D(h_{[CLS]}^{(x_s)}))$ and $\mathcal{L}_{ce}$ is the cross-entropy loss. Finally, the discriminator minimizes the following loss function:

$$\mathcal{L}_D = \mathcal{L}_{adv}(x_s, y_s, x_t) + \mathcal{L}_{reg}(x_s, y_s) \tag{5}$$

3.3 Classifier and generator losses

Following the principles of pseudo-labeling methods for domain adaptation, the ground truth label y_t for the target domain can be substituted by:

$$\hat{y}_t = \operatorname*{argmax}_k p_t^k \text{ if } p_t^k > \delta$$

where δ is a threshold. By using the learned confusion matrix $\eta^{(x_t)}$ to correct the pseudo-label generated by the classifier C, ALDA approximates the loss in the target domain by:

$$\mathcal{L}_T(x_t, \mathcal{L}_{unh}) = \sum_k c_k^{(x_t)} \mathcal{L}_{unh}(p_t, k) \tag{6}$$

where $\mathcal{L}_{unh}(p, k) = 1 - p_k$ is the unhinged loss. Then, the classifier C minimizes the following loss:

$$\mathcal{L}_C = \mathcal{L}_{ce}(p_s, y_s) + \lambda \, \mathcal{L}_T(x_t, \mathcal{L}_{unh}) \tag{7}$$

where $\mathcal{L}_{ce}(p_s, y_s)$ is the cross-entropy loss of the source domain. Finally, the generator G minimizes the following loss function:

$$\begin{aligned} \mathcal{L}_G = \mathcal{L}_{ce}(p_s, y_s) &+ \lambda \, \mathcal{L}_T(x_t, \mathcal{L}_{unh}) \\ &- \lambda \, \mathcal{L}_{adv}(x_s, y_s, x_t) \end{aligned} \tag{8}$$

where $\lambda \in [0, 1]$ is a hyperparameter.

4 Experiments

In this section, we present the experiments carried out to investigate the performance of our proposed method for Arabic cross-domain and cross-dialect sentiment analysis. We describe the used datasets and present the compared methods as well as the obtained results. We provide the experiments settings and implementation details of our method in Section A. The source code for reproducing the experimentations can be found in our github repository[1].

4.1 Datasets

We conduct three main sets of experiments to cover three possible scenarios.

Scenario 1: Domain adaptation for dialects of the same region. The set of experiments of this scenario aims to study our method's performance for cross-dialect and cross-domain sentiment analysis for Arabic dialects of the same region. To do so, we employ the existing multi-domain multi-dialect ArSentD-LEV dataset of the Levant region (Baly et al., 2018). ArSentD-LEV contains 1,000 tweets from each country of the Levant region (4,000 in total): Syria, Palestine, Jordan, and Lebanon. It is labeled into five classes and covers tweets from five topics: Personal (36%), Politics (23%), Religion (11%), Sport (6%), and Other (24%).

Scenario 2: Domain adaptation across regional dialects. In the set of experiments of this scenario, we investigate the performance of our method using the coarse-grained regional taxonomy of Arabic dialects. For this purpose,

[1] https://github.com/4mekki4/arabic-nlp-da

1. Firstly, we select three datasets, mixing Arabic dialects and MSA: BRAD (Elnagar and Einea), HARD (Elnagar et al., 2018), and TEAD (Abdellaoui and Zrigui, 2018) that are compiled from book reviews, hotel reviews, and Twitter, respectively. These datasets have sufficient samples to build a multi-dialect multi-domain dataset.

2. Secondly, we train an AraBERT-based dialect identification model, selecting data from some of the publicly available datasets, including MADAR (Bouamor et al., 2019), DART (Alsarsour et al., 2018), AOC (Zaidan and Callison-Burch, 2011), PADIC (Karima et al., 2018), and the multi dialect Arabic texts corpora proposed in (Khalid and Mark, 2013). The resulting Arabic dialect identification corpus consists of $353,171$ training sentences and a balanced test set of $50,000$ sentences and covers MSA as well as dialectal sentences from Maghrebi, Levantine, Egyptian, and Gulf. It is worth mentioning that our trained dialect identification model achieves 89% accuracy.

3. Finally, we apply our dialect identification model on the three evaluated datasets to build our multi-dialect multi-domain corpus. Moreover, we select the Levantine and Gulf dialects and MSA, which yielded sufficient data across domains. For review datasets, the rating levels 1 and 2 are assigned to negative polarity, while ratings 4 and 5 are considered positives. Furthermore, we sample 1000 positive and 1000 negative instances for each dialect to build our final multi-dialect and multi-domain dataset.

Scenario 3: Domain adaptation from MSA to Arabic dialects using social media data. The set of experiments of this scenario tackles the transfer of learning from MSA to Arabic dialects, belonging to different regions, using corpora built from social media (see Table 7 of Section 1.5 for the datasets details). Thus,

1. For MSA, we use the ArSAS dataset (Elmadany et al., 2018).

2. For the Maghrebi region, we employ MSAC (Morocco) (Oussous et al., 2020) and TSAC (Tunisia) (Medhaffar et al., 2017) datasets. Note that, we have removed sentences that are written in Arabezi for TSAC.

3. For the Egyptian region, we use the ASTD dataset (Nabil et al., 2015).

4. For the Levant region, we utilize AJGT (Jordan) (Alomari et al., 2017) and TweetSYR (Syria) (Salameh et al., 2015) datasets.

5. For the Gulf region, we employ the Saoudi dialect AraSenti-Tweet dataset (Al-Twairesh et al., 2017).

Since some of these datasets are labeled using positive and negative classes only (TSAC and MSAC), we evaluate our method using positive and negative sentences for all the used datasets.

We use the train-test splits of the evaluated datasets whenever this information is available. Otherwise, we split the datasets into 80% train and 20% test. For the ArSentD-LEV and following the work of (Khaddaj et al., 2019), we evaluate our method on the full target domain/dialect dataset. For all our experiments, we utilize the **accuracy** evaluation measure and highlight the best accuracy performance using bold font.

4.2 Compared Methods

In order to assess the performance of our method, we compare it with the state-of-the-art domain adaptation method, introduced by (Khaddaj et al., 2019), for Arabic sentiment analysis on the ArSentD-LEV dataset. Moreover, we evaluate BERT for zero-shot transfer from the source domain, denoted **ZS-BERT**. For a fair comparison, we investigate the performance of three state-of-the-art domain adaptation methods including **MMD** (Gretton et al., 2012), **CORAL** (Sun and Saenko, 2016), and **DANN** (Ganin et al., 2016). We implement the latter methods on top of AraBERT. We have also evaluated two state-of-the-art cross-domain sentiment analysis methods, namely **PBLM** (Ziser and Reichart, 2018) and **HTAN** (Li et al., 2018). It is worth to mention that for PBLM and HATN, we have used an extra 4000 unlabeled sentences from each domain/dialect. For HTAN, we have used Mazjak word embedding model (Abu Farha and Magdy, 2019)

4.3 Results

Scenario 1: Domain adaptation for dialects of the same region. Tables 1 and 2 present the results obtained for Arabic cross-domain and cross-dialect sentiment Analysis using ArSentD-LEV.

The overall obtained results for cross-dialect sentiment analysis (Table 1) show that ZS-BERT, the

		SOTA		Our Results				
Source	**Target**	**DANN$_{BOW}$**	**ADRL**	**ZS-BERT**	**CORAL**	**MMD**	**DANN**	**Ours**
Jordan	Lebanon	29	30	47	50	50.9	49.3	**52**
	Palestine	34.5	35	47.5	50.3	51.1	51.2	**52.8**
	Syria	32	33	51.7	53.3	53.2	51.9	**54.2**
Lebanon	Jordan	29	32	45	46.8	47.1	47.4	**48.8**
	Palestine	31	35	42.7	50.5	50.7	51	**52.4**
	Syria	37	37.5	49.6	50.7	51.1	50	**52**
Palestine	Jordan	32	32.5	45	50.6	49.7	47.4	**52.4**
	Lebanon	31	31	42	50	50.5	50.5	**51.9**
	Syria	28.5	27.5	51.7	52.4	52.4	51.3	**53.7**
Syria	Jordan	30.5	32	44.7	48.5	49.1	49.4	**51**
	Lebanon	35	35.5	46.1	51.5	51.1	50.6	**52**
	Palestine	31.5	37.5	47.1	49.7	49.8	51.3	**52.9**

Table 1: The results of accuracy measurement of Arabic cross-dialect sentiment analysis using the ArSentD-LEV dataset. The SOTA results are taken from (Khaddaj et al., 2019).

zero-shot transfer-based method, outperforms both DANN$_{BOW}$ and ADRL, the state-of-the-art domain adaptation methods that are based on the bag-of-words representation. Moreover,training the state-of-the-art domain adaptation methods, including CORAL, MMD and DANN, on top of BERT module has improved BERT transfer performance across dialects. Besides, these three methods achieve comparable performance for most source and target dialects and outperform both DANN$_{BOW}$ and ADRL. Furthermore, our method, which is based on BERT and ALDA's losses, surpasses the existing state-of-the-art methods and ZS-BERT with average improvements of 19% and 5.5% respectively. Additionally, it shows better performance than the other domain adaptation methods that are implemented on top of BERT (CORAL, MMD, and DANN).

In accordance with the results obtained for cross-dialect, Table 2 shows that the ZS-BERT method outperforms both DANN$_{BOW}$ and ADRL in most test cases of cross-domain sentiment analysis (14 out of 20 cases). Besides, the results show that the three domain adaptation methods CORAL, MMD, and DANN outperform both DANN$_{BOW}$ and ADRL, and improve the transfer performance of BERT model. On average, the latter three methods (CORAL, MMD, and DANN) are on a par with each other in terms of accuracy. Similarly, our proposed method outperforms both state-of-the-art methods (DANN$_{BOW}$ and ADRL) as well as ZS-BERT by an average increment of 19% and 10.7%, respectively. Moreover, it achieves a better performance than CORAL, MMD, and DANN for most source and target domains (12 out of 20 cases).

		SOTA		Our Results				
Source	**Target**	**DANN$_{BOW}$**	**ADRL**	**ZS-BERT**	**CORAL**	**MMD**	**DANN**	**Ours**
Politics	Personal	28.7	33.3	28.7	41.6	41.3	43	**44.3**
	Religious	20.3	25.3	10	33.6	33.3	34.2	**46.3**
	Sport	35.1	35.1	36.7	46.6	32.8	**46.8**	**46.8**
	Other	22.5	24.2	38.2	**49.7**	50	39.7	46.1
Personal	Politics	41.7	36.8	46.3	49.7	49.4	47.5	**49.7**
	Religious	22.8	23.4	41	44.3	**44.7**	43.5	44.2
	sport	26.8	25.8	43.5	**49.7**	49.5	48.2	46.6
	Other	33.8	35.4	53	57.4	57.7	49.6	**58**
Religious	Politics	15.5	15.5	12	**42**	42	37.6	40.8
	Personal	24.1	26.1	25	35.1	37	36.8	**38**
	Sport	25.8	26.8	21.6	**38.1**	32.8	28.5	34.8
	Other	30.6	27.4	26.4	46.4	43.2	43.2	**48.4**
Sport	Politics	36.4	30.7	46.9	**48.7**	48.3	43.1	44.6
	Personal	25.3	24.5	40.7	43.8	42.3	43.6	**44.5**
	Religious	20	19	30.8	29.2	31	40.2	**44**
	Other	35.5	35.5	48.3	49	49.6	49	**54.2**
Other	Politics	23.2	23.2	**46.8**	46.5	46.4	34.4	**46.8**
	Personal	30.3	24.9	40.2	**46.2**	44.3	40.3	45.5
	Religious	41.8	43	39.5	45.8	47.6	48.6	**48.9**
	Sport	23.7	27.8	46.7	48.4	**51.1**	47.7	50.9

Table 2: The results of accuracy measurement of Arabic cross-domain sentiment analysis using the ArSentD-LEV dataset. The SOTA results are taken from (Khaddaj et al., 2019).

Scenario 2: Domain adaptation across regional dialects. Table 3 summarizes the results obtained for cross-domain and cross-dialect as well as cross-domain and cross-dialect Arabic sentiment analysis using two regional dialects (Gulf and Levantine) and MSA data, covering three domains (books reviews, hotels reviews and Twitter).

The overall obtained results show that the zero-shot transfer from AraBERT (ZS-BERT) outperforms previous state-of-the-art methods (PBLM and HTAN). Moreover, the evaluated domain adaptation methods on top of BERT improve AraBERT's performance for all evaluated scenarios. Besides, the results demonstrate that the performance of ZS-BERT method drops significantly in the cases of cross-domain as well as in cross-domain and cross-dialect scenarios. Nevertheless, the domain adaptation methods show more important improvements (an increment of 7.4% on average) in the scenarios mentioned above. The obtained results clearly show that our method surpasses the other methods for most target datasets and scenarios, except for some cases but the gap remains small.

Scenario 3: Domain adaptation from MSA to Arabic dialects using social media data. Table 4 presents the domain adaptation results obtained

Scenario	Target data	Gulf				Levantine				Modern Standard Arabic			
		BRAD	**HARD**	**TEAD**	**Avg**	**BRAD**	**HARD**	**TEAD**	**Avg**	**BRAD**	**HARD**	**TEAD**	**Avg**
Cross-dialect	**PBLM**	53.5	83.38	66.37	67.75	57.25	78.12	62.62	66.0	50.87	80.0	64.75	65.21
	HATN	58.05	75.5	62.2	65.25	58.75	74.6	61.72	65.02	58.22	77.35	57.98	64.52
	ZS-BERT	72.7	92.1	72.5	79.1	80.1	95.3	73.8	83.1	75.3	95.3	73.0	81.2
	CORAL	77.0	94.1	**73.3**	81.5	81.1	95.4	73.6	83.4	**80.3**	96.6	75.5	84.1
	MMD	77.1	**94.4**	73.1	81.5	81.9	96.3	74.1	84.1	80.1	**97.1**	**76.8**	**84.7**
	DANN	77.6	94.3	72.9	81.6	81.4	95.8	75.6	84.3	79.1	96.4	74.5	83.3
	Ours	**78.4**	94.3	73.0	**81.9**	**82.3**	**96.5**	**76.6**	**85.1**	79.6	96.4	74.9	83.6
Cross-domain	**PBLM**	51.25	51.63	48.88	50.58	59.62	50.25	48.25	52.71	64.0	51.75	50.0	55.25
	HATN	57.35	54.45	52.6	54.8	52.77	49.32	48.88	50.32	60.7	52.97	54.68	56.12
	ZS-BERT	55.7	70.9	58.8	61.8	60.6	69.9	58.0	62.8	64.5	76.8	61.8	67.7
	CORAL	62.9	82.3	60.6	68.6	**66.1**	74.1	59.9	66.7	66.3	78.3	**66.9**	70.5
	MMD	62.3	73.0	61.5	65.6	64.4	75.8	59.4	66.5	67.4	82.6	66.1	72.0
	DANN	62.9	80.1	59.8	67.6	62.4	77.1	**62.5**	67.3	66.6	78.6	66.4	70.5
	Ours	**65.3**	**85.1**	**62.5**	**71.0**	64.5	**79.1**	62.3	**68.6**	**69.8**	**93.3**	66.8	**76.6**
Cross-dialect & Cross-domain	**PBLM**	51.56	50.62	49.81	50.67	50.0	52.44	48.81	50.42	53.19	49.44	49.56	50.73
	HATN	55.24	50.91	52.29	52.81	56.78	50.24	50.2	52.4	55.35	52.36	53.06	53.59
	ZS-BERT	57.8	72.9	59.9	63.5	63.0	74.5	59.1	65.5	60.2	71.6	60.8	64.2
	CORAL	63.8	82.8	60.8	69.1	**64.8**	78.3	60.8	67.9	**66.8**	79.4	63.9	70.0
	MMD	63.6	82.0	60.5	68.7	64.4	79.4	60.9	68.3	66.4	77.9	64.1	69.5
	DANN	63.2	75.8	**62.3**	67.1	63.7	77.2	**62.1**	67.7	65.2	**80.2**	65.2	70.2
	Ours	**64.9**	**85.6**	61.7	**70.8**	64.6	**86.5**	61.2	**70.8**	**66.8**	79.4	**65.8**	**70.6**

Table 3: The results of accuracy measurement of cross-dialect and cross-domain as well as cross-domain & cross-dialect Arabic sentiment analysis using two regional dialects and MSA data, covering three domains (books, hotels, and Twitter). Each target dataset's performance is the average accuracy obtained using its corresponding domain and/or dialect source data for each scenario. For example, in the cross-dialect scenario, the result of Gulf_BRAD is the average accuracy obtained from Levantine_BRAD and MSA_BRAD as source dialect.

from MSA to Arabic dialects. In agreement with the previously obtained results, all domain adaptation methods outperform the ZS-BERT method for all evaluated datasets by an average increment of 4.9% . CORAL, MMD, and DANN achieve comparable performances for most dialectal datasets. Moreover, the overall comparison results show that our method outperforms all other domain adaptation methods.

Target	**ZS-BERT**	**CORAL**	**MMD**	**DANN**	**Ours**
MSAC	85	88	88.2	87.7	**89.5**
TSAC	81.4	84.9	84	86.2	**87.5**
ASTD	87.3	91.1	90.2	90.2	**91.5**
AJGT	83.8	90	88.8	85.5	**90.5**
TweetSYR	83.8	85.8	86.9	84.7	**87.5**
AraSenti-Tweet	79.6	80.4	80.7	81.5	**83.9**

Table 4: The results of accuracy measurement of domain adaptation results from MSA (ArSAS source dataset) to Arabic dialects.

4.4 Result discussion

The overall obtained results of the evaluated scenarios show that our method improves the transfer performance from contextualized word embedding. Moreover, it achieves far better transfer performance than the state-of-the-art methods that are based on the bag-of-words representation or pretrained word embedding. Indeed, all BERT-based domain adaptation methods yield a far better transfer learning performance than both DANN_BOW and ADRL methods. Besides, our method achieves better performance than CORAL, MMD, and DANN, which are implemented on top of BERT module. These results can be explained by the fact that BERT captures a high-level representation of the input text (Devlin et al., 2019; Antoun et al., 2020) as well as the effectiveness of ALDA. In fact, the latter aligns both domain output distributions using adversarial training and captures the discriminative features of the target domain inputs throughout self-training (Chen et al., 2020). Moreover, using BERT as a feature generator allows the model to extract high-level shared features of the input data that are transferable across domains and dialects. For instance, the results show that training DANN on top of BERT model outperforms DANN$_{BOW}$, trained using the bag-of-words text representation, or even state-of-the-art methods that are based on pivot features extraction (HATN and PBLM), by a large margin for both cross-domain and cross-dialect sentiment analysis (Table 2 and

Table 1).

5 Error Analysis

To understand why our proposed method outperforms the previous methods, we perform an error analysis. In this error analysis we focus on two aspects: the misclassified instances by our system and the instances correctly predicted by our method which the other approaches fail.

For the first aspect, the majority of misclassified samples correspond to very short sentences in the target dialect. Most of them are either idiomatic, offensive or sarcastic expressions that are specific to the target dialect and contains words that are distant from MSA :واعرة /wAErp/, غار الله يعفو وصافي /gAr Allh yEfw wSAfy/, ملا طحان /mlA THAn/ and شرف خارف /crf xArf/[2]. It is worth mentionning that the other evaluated methods also misclassify these samples.

For the second aspect, we have checked the cases where our method correctly predicts the instances labels while the other methods fail. Overall, we notice that the zero-shot predictions were biased to the distribution of the source data, as example the ArSAS dataset contains 63% of negative instances. MMD, CORAL and DANN overcome this issue by aligning the distribution of source and target features, which improves the results on the target domain. Meanwhile, they tend to misclassify reviews that convey multiple sentiment polarities, as the case for hotel reviews or books reviews, where users tend to express their negative and positive sentiments in the same review. Table 8 (Section B) shows a sample of these instances. Our method outperforms these DA methods since it relies on a noise-correcting discriminator that generates different noise vectors for the source and the target domain and learns a confusion matrix in an adversarial manner. By correcting the noise in pseudo labels of the target domain using the confusion matrix, we can achieve a class-wise feature alignment of the source and the target domains. Nevertheless, the other evaluated DA methods align the output features of the source and the target domain in class agnostic fashion.

6 Conclusion

In this work, we have introduced an unsupervised domain adaptation method for Arabic cross-domain and cross-dialect sentiment analysis based on the pretrained AraBERT language model and the Adversarial-Learned Loss for Domain Adaptation (ALDA). We have performed several experiments to investigate the performance of our method as well as several state-of-the-art methods, adopting both the coarse-grained and the fine-grained taxonomies of Arabic dialects. Moreover, we have studied the performance of domain adaptation from the MSA to Arabic dialects using social media data. The overall obtained results showed that domain adaptation methods outperform zero-shot transfer from BERT model by a large margin. Furthermore, our method achieved a very promising performance and surpassed the evaluated methods on most test datasets.

In future work, we plan to investigate domain adaptive pre-training by collecting unlabeled data for target domains and fine-tuning AraBERT using the MLM objective. The aim is to study the performance of our method using domain aware language model. Since the zero-shot transfer performance using BERT model drops significantly in cross-domain sentiment analysis experiments, we believe that training domain adaptation methods on top of domain aware BERT model will lead to improved performance. We also plan to study domain adaptation from rich-resource languages such as English to Arabic language and its dialects.

References

Houssem Abdellaoui and Mounir Zrigui. 2018. Using Tweets and Emojis to Build TEAD: an Arabic Dataset for Sentiment Analysis. Computación y Sistemas, 22:777 – 786.

Muhammad Abdul-Mageed, Chiyu Zhang, Azadeh Hashemi, and El Moatez Billah Nagoudi. 2020. AraNet: A deep learning toolkit for Arabic social media. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, pages 16–23, Marseille, France. European Language Resource Association.

Ibrahim Abu Farha and Walid Magdy. 2019. Mazajak: An online Arabic sentiment analyser. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 192–198, Florence, Italy. Association for Computational Linguistics.

Mahmoud Al-Ayyoub, Abed Allah Khamaiseh, Yaser Jararweh, and Mohammed N. Al-Kabi. 2019. A comprehensive survey of arabic sentiment analysis. Information Processing & Management, 56(2):320 –

[2] Transliteration is performed using Safe Buckwlater scheme

342. Advance Arabic Natural Language Processing (ANLP) and its Applications.

Ahmad Al Sallab, Hazem Hajj, Gilbert Badaro, Ramy Baly, Wassim El Hajj, and Khaled Bashir Shaban. 2015. Deep learning models for sentiment analysis in Arabic. In Proceedings of the Second Workshop on Arabic Natural Language Processing, pages 9–17, Beijing, China. Association for Computational Linguistics.

Nora Al-Twairesh, Hend Al-Khalifa, AbdulMalik Al-Salman, and Yousef Al-Ohali. 2017. Arasenti-tweet: A corpus for arabic sentiment analysis of saudi tweets. Procedia Computer Science, 117:63 – 72. Arabic Computational Linguistics.

Khaled Mohammad Alomari, Hatem M. ElSherif, and Khaled Shaalan. 2017. Arabic tweets sentimental analysis using machine learning. In Advances in Artificial Intelligence: From Theory to Practice, pages 602–610, Cham. Springer International Publishing.

Israa Alsarsour, Esraa Mohamed, Reem Suwaileh, and Tamer Elsayed. 2018. DART: A large dataset of dialectal Arabic tweets. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Mohamed Aly and Amir Atiya. 2013. LABR: A large scale Arabic book reviews dataset. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 494–498, Sofia, Bulgaria. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, pages 9–15, Marseille, France. European Language Resource Association.

Elnagar Ashraf and Einea Omar. 2016. Brad 1.0: Book reviews in arabic dataset. In 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), pages 1–8.

Gilbert Badaro, Ramy Baly, Hazem Hajj, Wassim El-Hajj, Khaled Bashir Shaban, Nizar Habash, Ahmad Al-Sallab, and Ali Hamdi. 2019. A survey of opinion mining in arabic: A comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations. ACM Trans. Asian Low-Resour. Lang. Inf. Process., 18(3).

Gilbert Badaro, Ramy Baly, Hazem Hajj, Nizar Habash, and Wassim El-Hajj. 2014. A large scale Arabic sentiment lexicon for Arabic opinion mining. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), pages 165–173, Doha, Qatar. Association for Computational Linguistics.

Ramy Baly, Alaa Khaddaj, Hazem M. Hajj, Wassim El-Hajj, and Khaled Bashir Shaban. 2018. ArSentD-LEV: A multi-topic corpus for target-based sentiment analysis in arabic levantine tweets. In OSACT 3: The 3rd Workshop on Open-Source Arabic Corpora and Processing Tools, page 37.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 199–207, Florence, Italy. Association for Computational Linguistics.

Minghao Chen, Shuai Zhao, Haifeng Liu, and Deng Cai. 2020. Adversarial-learned loss for domain adaptation. In The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 3521–3528. AAAI Press.

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.

Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. 2020. Adversarial and domain-aware BERT for cross-domain sentiment analysis. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4019–4028, Online. Association for Computational Linguistics.

Samhaa R. El-Beltagy. 2016. NileULex: A phrase and word level sentiment lexicon for Egyptian and Modern Standard Arabic. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 2900–2905, Portorož, Slovenia. European Language Resources Association (ELRA).

AbdelRahim Elmadany, Hamdy Mubarak, and Walid Magdy. 2018. Arsas: An arabic speech-act and sentiment corpus of tweets. Proceedings of the 3rd Workshop on Open-source Arabic Corpora and Processing Tools OSACT'2018, 3:20.

Ashraf Elnagar and Omar Einea. Brad 1.0: Book reviews in arabic dataset. In 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), pages 1–8. IEEE.

Ashraf Elnagar, Yasmin S. Khalifa, and Anas Einea. 2018. Hotel Arabic-Reviews Dataset Construction for Sentiment Analysis Applications, pages 35–52. Springer International Publishing, Cham.

Hady ElSahar and Samhaa R. El-Beltagy. 2015. Building large arabic multi-domain resources for sentiment analysis. In Computational Linguistics and Intelligent Text Processing, pages 23–34, Cham. Springer International Publishing.

Alexander Erdmann, Nasser Zalmout, and Nizar Habash. 2018. Addressing noise in multidialectal word embeddings. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers, pages 558–565. Association for Computational Linguistics.

Fadhl Eryani, Nizar Habash, Houda Bouamor, and Salam Khalifa. 2020. A spelling correction corpus for multiple arabic dialects. In Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020, pages 4130–4138. European Language Resources Association.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. J. Mach. Learn. Res., 17(1):2096–2030.

Hazem Hajj Wassim El-Hajj Gilbert Badaro, Hussein Jundi and Nizar Habash. 2018. Arsel: A large scale arabic sentiment and emotion lexicon. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Paris, France. European Language Resources Association (ELRA).

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. Journal of Machine Learning Research, 13(25):723–773.

Nizar Habash, Fadhl Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, Sara Hassan, Faisal Al-Shargi, Sakhar B. Alkhereyfy, Basma Abdulkareem, Ramy Eskander, Mohammad Salameh, and Hind Saddiki. 2018. Unified guidelines and resources for arabic dialect orthography. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Nizar Y Habash. 2010. Introduction to Arabic natural language processing, volume 3. Morgan & Claypool Publishers.

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.

Meftouh Karima, Harrat Salima, and Smaïli Kamel. 2018. PADIC: extension and new experiments. In 7th International Conference on Advanced Technologies ICAT, Antalya, Turkey.

Alaa Khaddaj, Hazem Hajj, and Wassim El-Hajj. 2019. Improved generalization of Arabic text classifiers. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 167–174, Florence, Italy. Association for Computational Linguistics.

Almeman Khalid and Lee Mark. 2013. Automatic building of arabic multi dialect text corpora by bootstrapping dialect words. In 2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA), pages 1–6.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. Advances in Neural Information Processing Systems (NeurIPS).

Juntao Li, Ruidan He, Hai Ye, Hwee Tou Ng, Lidong Bing, and Rui Yan. 2020. Unsupervised domain adaptation of a pretrained cross-lingual language model. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, pages 3672–3678. International Joint Conferences on Artificial Intelligence Organization. Main track.

Zheng Li, Ying Wei, Yu Zhang, and Qiang Yang. 2018. Hierarchical attention transfer network for cross-domain sentiment classification.

Salima Medhaffar, Fethi Bougares, Yannick Estève, and Lamia Hadrich-Belguith. 2017. Sentiment analysis of Tunisian dialects: Linguistic ressources and experiments. In Proceedings of the Third Arabic Natural Language Processing Workshop, pages 55–61, Valencia, Spain. Association for Computational Linguistics.

Leila Moudjari, Karima Akli-Astouati, and Farah Benamara. 2020. An Algerian corpus and an annotation platform for opinion and emotion analysis. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 1202–1210, Marseille, France. European Language Resources Association.

Abdul-Mageed Muhammad, Zhang Chiyu, Bouamor Houda, and Habash Nizar. 2020. NADI 2020: The first nuanced arabic dialect identification shared task. arXiv:2010.11334.

Mahmoud Nabil, Mohamed Aly, and Amir Atiya. 2015. ASTD: Arabic sentiment tweets dataset. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2515–2519, Lisbon, Portugal. Association for Computational Linguistics.

Oumaima Oueslati, Erik Cambria, Moez Ben HajHmida, and Habib Ounelli. 2020. A review of sentiment analysis research in arabic language. Future Generation Computer Systems, 112:408 – 430.

Ahmed Oussous, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, and Samir Belfkih. 2020. Asa: A framework for arabic sentiment analysis. Journal of Information Science, 46(4):544–559.

Chatrine Qwaider, Stergios Chatzikyriakidis, and Simon Dobnik. 2019. Can Modern Standard Arabic approaches be used for Arabic dialects? sentiment analysis as a case study. In Proceedings of the 3rd Workshop on Arabic Corpus Linguistics, pages 40–50, Cardiff, United Kingdom. Association for Computational Linguistics.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in nlp—a survey. arXiv preprint arXiv:2006.00632.

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. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 4933–4941, Marseille, France. European Language Resources Association.

Houda Saadane and Nizar Habash. 2015. A conventional orthography for algerian arabic. In Proceedings of the Second Workshop on Arabic Natural Language Processing, ANLP@ACL 2015, Beijing, China, July 30, 2015, pages 69–79. Association for Computational Linguistics.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1332–1344, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mohammad Salameh, Saif Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on Arabic social media posts. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 767–777, Denver, Colorado. Association for Computational Linguistics.

Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In Computer Vision – ECCV 2016 Workshops, pages 443–450, Cham. Springer International Publishing.

Thuy-Trang Vu, Dinh Phung, and Gholamreza Haffari. 2020. Effective unsupervised domain adaptation with adversarially trained language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, pages 6163–6173. Association for Computational Linguistics.

Hai Ye, Qingyu Tan, Ruidan He, Juntao Li, Hwee Tou Ng, and Lidong Bing. 2020. Feature adaptation of pre-trained language models across languages and domains with robust self-training. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, pages 7386–7399. Association for Computational Linguistics.

Jianfei Yu and Jing Jiang. 2016. Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 236–246, Austin, Texas. Association for Computational Linguistics.

Omar F. Zaidan and Chris Callison-Burch. 2011. The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 37–41, Portland, Oregon, USA. Association for Computational Linguistics.

Omar F. Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. Computational Linguistics, 40(1):171–202.

Fuzhen Zhuang, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, and Qing He. 2015. Supervised representation learning: Transfer learning with deep autoencoders. In Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI'15, page 4119–4125. AAAI Press.

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.

A Experiments settings

1.1 Implementation details

We implement our method as well as the other evaluated methods using PyTorch. For all our experiments, AraBERT is used as the input text representation generator. The sentiment classification model is a fully-connected layer that takes the input text representation generated by AraBERT and outputs label probabilities through the Softmax function. For hyperparameters tuning, we have followed the method of Chen et al. (2020). For tuning the learning rate, we have conducted a random search over the set of values $\{310^{-5}, 210^{-5}, 10^{-5}, 810^{-6}, 510^{-6}\}$. We employ Adam optimizer with a learning rate of 510^{-6} for our method and DANN, while a learning rate of 10^{-5} is adopted for the CORAL and MMD and 210^{-5} for ZS-BERT. We use the same discriminator architecture as in (Chen et al., 2020). The learning rate of the discriminator is set to be ten times the value of the generator. During the training, the learning is adjusted at every iteration using $\eta_p = \frac{\eta_0}{(1+\alpha \cdot q)^{\beta}}$ (Chen et al., 2020), where q is the training progress, η_0 = 0.01, $\alpha = 10$ and $\beta = 0.75$. The hyperparameter λ is varied between 0 and 1 during the training using $\frac{2}{1+exp(10 \cdot q)} - 1$ (Chen et al., 2020). The self-labeling threshold (varied between 0.6 and 0.9) is fixed to 0.8, except for the ArSentD-LEV dataset, where it is fixed to 0.6. For all the evaluated methods, we use a batch-size of 16 samples and 20 training epochs, except for ZS-BERT, where the number of epochs is fixed at 5. Table 5 presents the number of trainable parameters for each domain adaptation method.

Method	Genrator	Classifier	Discriminator	Total
ZS-BERT	135197189	3845	−−	135201034
CORAL	135197189	3845	−−	135201034
MMD	135197189	3845	−−	135201034
DANN	135197189	3845	1181953	136382987
Ours	135197189	3845	1185029	136386063

Table 5: The number of trainable parameters for each domain adaptation method in scenario 1 (number of output classes equals to 5).

1.2 Computing Infrastructure

We conduct our experiments using an Intel(R) Xeon(R) Gold 6152 CPU @ 2.10GHz working station, having a single Nvidia Tesla P100 with 16GB of RAM.

1.3 Average Runtimes

Tabe 6 presents the average runtime of a single run of domain adaptation methods and ZS-BERT (for the cross-domain and cross-dialect experiments).

Method	Average runtime
ZS-BERT	2m30s
MMD	11m26s
CORAL	14m20s
DANN	14m46s
Ours	21m50s

Table 6: Average runtime of each method for cross-domain and cross-dialect experiments (Table 3).

1.4 Evaluation measure

For all our experiments, we have employed the accuracy evaluation measure:

$$accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

1.5 Details of the used datasets in Scenario 3

Table 7 shows the details of the train/test splits used in scenario 3 (Section 4.1). We use 20% of the dataset as a test set, except for the ArSenti-Tweet dataset where the train-test is available.

Dataset	Size	Train	Test
ArSAS	11109	80%	20%
MSAC	2000	80%	20%
TSAC	5506	80%	20%
ASTD	1589	80%	20%
AJGT	1800	80%	20%
TweetSYR	1798	80%	20%
ArSenti-Tweet	11112	9750	1362

Table 7: The description of the datasets used in Scenario 3. When the train-test split is not available, we use 20% of the data for the test set.

1.6 Datasets links

- Scenario 1: ArSentD-LEV
- Scenario 2: BRAD, HARD, TEAD
- Scenario 3: TSAC, ASTD, AJGT, ArSAS MSAC, TweetSYR, AraSenti-Tweet[3].

B Error analysis

Table 8 shows a sample of sentences along with their predicted label by our proposed model, ZS-BERT, and DANN.

[3] The AraSenti-Tweet dataset is delivered by the authors (Al-Twairesh et al., 2017)

Sentence	Ground-truth	ZS-BERT	DANN	Ours
استثنائي النظافة والبهو واستقبال ضعف اتصال شبكة موبايلي /AstvnAQy AlnZAfp wAlbhw wAstqbAl DEf AtSAl cbkp mwbAyly/	Positive	Negative	Negative	Positive
ضعيف بوفية الافطار الرمضاني ممتاز عدم الالتزام بموعد الدخول سوء معاملة موظفين الاستقبال وقاحة مدير موظفين الاستقبال /DEyf bwfyp AlAfTAr AlrmDAny mmtAz Edm AlAltzAm bmwEd Aldxwl swC mEAmlp mwZfyn AlAstqbAl wqAHp mdyr mwZfyn AlAstqbAl/	Negative	Negative	Positive	Negative
جيد الموقع النظافه الحراسه الخصوصيه السعر مبالغ فيهاالإنترنتثلاجه الغرفه لا يوجد بها شئ منع دخول المأكولات والمشروبات اسعار المطعم مبالغ فيها بشكل خيالي /jyd AlmwqE AlnZAfh AlHrAsh AlxSwSyh AlsEr mbAlg fyhAlIntrntvlAjh Algrfh lA ywjd bhA cQ mnE dxwl AlmOkwlAt wAlmcrwbAt AsEAr AlmTEm mbAlg fyhA bckl xyAly/	Positive	Negative	Negative	Positive
جيد موظفي الاستقبال بطئ في التعامل وعدم خبره كافيه /jyd mwZfy AlAstqbAl bTQ fy AltEAml wEdm xbrh kAfyh/	Positive	Negative	Negative	Negative

Table 8: Examples of predictions made by our proposed system compared with the Zero-Shot BERT and DANN