Improved Generalization of Arabic Text Classifiers

Alaa Khaddaj Hazem Hajj Wassim El-Hajj
American University of Beirut
Beirut, Lebanon
{awk11, hh63, we07}@aub.edu.lb

Abstract

While transfer learning for text has been very active in the English language, progress in Arabic has been slow, including the use of Domain Adaptation (DA). Domain Adaptation is used to generalize the performance of any classifier by trying to balance the classifier's accuracy for a particular task among different text domains. In this paper, we propose and evaluate two variants of a domain adaptation technique: the first is a base model called Domain Adversarial Neural Network (DANN), while the second is a variation that incorporates representational learning. Similar to previous approaches, we propose the use of proxy A-distance as a metric to assess the success of generalization. We make use of ArSentD-LEV, a multi-topic dataset collected from the Levantine countries, to test the performance of the models. We show the superiority of the proposed method in accuracy and robustness when dealing with the Arabic language.

1 Introduction

Natural Language Processing (NLP) for Arabic is challenging due to the complexity of the language. Additionally, resources in Arabic are scarce making it difficult to achieve NLP progress at the pace of other resource-rich languages such as English (Badaro et al., 2019). As a result, there is a need for transfer learning methods that can overcome the resource limitations. In this paper, we propose the use of domain adaptation to address this challenge while considering the task of sentiment analysis (SA) also referred to as Opinion Mining (OM).

When training over a dataset with multiple domains, different domains have different data distributions. This has a negative impact when training on one domain and testing on another, since the model would not be able to generalize well. Although domains within the same dataset have differences, they share some characteristics. For example, consider reviews of Amazon products: reviews of electronic products are different from book reviews, but these two domains share the general structure of reviews. We say there exists a shift in the data's distribution between the two domains. To solve this problem, many approaches were proposed within the field of Domain Adaptation (DA) (Ben-David et al., 2010). This field is receiving a lot of attention in English, a lot more than its Arabic counterpart.

Solving the data shift problem is of interest for many reasons. First, it is harder for machine learning to learn good internal representations on the Arabic text as opposed to English text. This is due to the sparsity of the Arabic language, and its morphological complexity compared to English. Another reason is the limited amount of available data, especially for dialects, which causes deep learning models to perform bad on any task. Lastly, we are not aware of domain adaptation techniques for the Arabic language, and thus much work needs to be done in this area to catch up with the research in English.

Traditionally, researchers focused their efforts on extracting features shared between the source and target domains (Blitzer et al., 2006, 2007; Pan et al., 2010). After the advancement of representational learning (Bengio et al., 2013), several algorithms were introduced. The most notable approaches are Stacked Denoising Autoencoder (SDA) (Vincent et al., 2010; Glorot et al., 2011). Later, a modified version was introduced by (Chen et al., 2012). This version, called marginalized Stacked Denoising Autoencoder (mSDA), introduced a speedup compared to the original SDA since the input/output relation was provided in closed form. After Generative Adversarial Nets (Goodfellow et al., 2014) were

Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 167–174
Florence, Italy, August 1, 2019.

introduced, the interest in adversarial training increased. Researchers developed new approaches that solve the DA problem through adversarial training, with emphasis on applications in computer vision and limited exploration for NLP. The most notable approaches are Domain Adversarial Neural Networks (DANN) (Ganin et al., 2016), Domain Separation Network (DSN) (Bousmalis et al., 2016), Adversarial Discriminative Domain Adaptation (ADDA) (Tzeng et al., 2017) and Conditional Adversarial Domain Adaptation (Long et al., 2018). Although limited in Arabic, some efforts have been spent to solve the domain shift problem (Jeblee et al., 2014; Monroe et al., 2014).

In this paper, we propose and evaluate some adversarial approaches for domain adaptation. The first is a regular DANN model while the second is a variant of DANN that incorporates representational learning. To assess the success of domain adaptation, we use the proxy A-distance as a matrix (Ben-David et al., 2007). The rest of the paper is organized as follows. Section 2 presents different approaches for DA. Section 3 introduces the algorithms to be evaluated, and describes the dataset. Section 4 presents the experiments and the results. We finally summarize our work and conclude the paper in Section 5.

2 Related Work

Domain Adaptation passed through several development stages. The first stage was based on feature engineering methods, while in the later stages, DA experienced a shift towards deep learning.

Initial approaches included finding words that behaved similarly in both the source and target domains. Blitzer et al. (2006) called such words *pivot features*, and proposed different approaches for extracting them. He first proposed using the most frequent common words as pivot features (Blitzer et al., 2006), and later on proposed using words with highest mutual information with the source labels (Blitzer et al., 2007). The extracted pivot features are then used by the algorithm to augment the initial dataset. This is done by learning a mapping to a vector space with dimensionality smaller than the dimensionality of the input data. Then, an optimization problem is solved in the new space, with the objective function being a similarity measure. Using the results of the optimization problem, new features are added to the original dataset. The resulting algorithm is called Structural Correspondence Learning (SCL) (Blitzer et al., 2006, 2007). A similar approach was introduced by Gong et al. (2013) where they suggested finding words, which they called *landmarks*, that have similar distributions over the source and target domains. These landmarks were used to increase the confusion between source and target domains, through optimizing a series of auxiliary tasks. Another point of view was introduced by Pan et al. (2010) based on the Spectral Graph Theory. Their approach, called Spectral Feature Alignment (SFA), aligned features from source and target domains using bipartite graphs. Although these approaches improved accuracies in domain adaptation tasks, the improvements remained limited.

The hype of deep learning motivated finding deep learning algorithms that could solve this problem. An interesting approach by Glorot et al. (2011) was preparing the input of any classifier by passing the input through Stacked Denoising Autoencoders (SDA) (Vincent et al., 2010). The use of SDAs helps find a new representation of the data that is domain invariant. This is achieved by reconstructing the input from stochastically disrupted data (via noise injection). Once the data is transformed, a linear SVM is trained on the new representation. This approach was more accurate than the previous approaches in predicting target domain labels. However, training SDAs is very time consuming. That is why Chen et al. (2012) forced the reconstruction mapping to be linear. This restriction yielded a closed form output solution. The new model, called marginalized Stacked Denoising Autoencoder (mSDA), was able to perform as good as the original SDA, and took much less time for training.

After the publication of GANs (Goodfellow et al., 2014), many researchers took interest in adversarial training. Ganin et al. (2016) proposed an adversarial network for domain adaptation. By introducing a Gradient Reversal Layer (GRL) that inverts the gradient's sign during backpropagation, the Domain Adversarial Neural Network (DANN) was forced to find a saddle point between 2 errors: a label prediction error (that is to be minimized) and a domain classification error (to be maximized). This approach led to the emergence of domain invariant features. DANN achieved state-of-the-art performance in domain adaptation tasks for two specific applications, namely: senti-

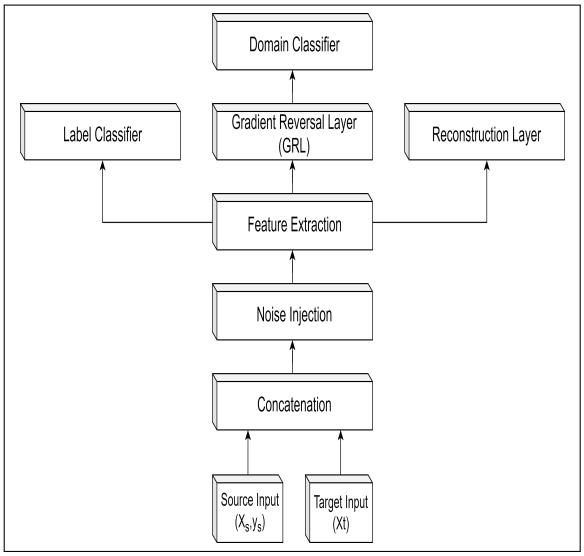

Figure 1: Proposed Model Architecture

ment analysis and computer vision.

For the Arabic language, the domain adaptation research area is still very limited. Joty et al. investigated the problem of cross-language adaptation for question-question similarity, and proposed a Cross-Language Adversarial Neural Network (CLANN) (Joty et al., 2017). Monroe et al. used *feature space augmentation* presented by (Daume III, 2007) for word segmentation (Monroe et al., 2014). Both approaches were successful.

3 Proposed Method

A Domain Adaptation task is, in general, a prediction problem where given label data from a source domain S, we are to predict the labels of a target domain T with unlabeled data (Ben-David et al., 2010). In this paper, we focus on domain adaptation for sentiment analysis: Given data with sentiment labels from one domain, the model should be able to predict the sentiment of data coming from another domain.

Let $(X_s, Y_s) = \{(x_i, y_i)\}_{i=1}^{N_s}$ represent the source domain input data of N_s observations x_i, where x_i could be any textual data (*e.g.* Bag-Of-Words, Sequence, etc...), and y_i the corresponding label. The domain input data $X_t = \{x_i\}_{i=1}^{N_t}$ consists of N_t unlabeled observations. The source and target observations are concatenated to form the input data X of $N_s + N_t$ observations to the model. The architecture of DANN adopted is similar to the one in (Ganin et al., 2016). The variant, shown in Figure 1, is composed of 5 main parts:

- Feature Extractor
- Label Predictor
- Reconstruction Layer
- Domain Predictor
- Gradient Reversal Layer

The above model uses denoising reconstruction (Vincent et al., 2010; Chen et al., 2012) and adversarial training (Ganin et al., 2016), in order to learn features that are discriminative towards the tasks at hand, while at the same time being able to generalize from one domain to another.

Three loss functions are associated with the network: 1) a loss function related to the classification task at hand, denoted as $\mathcal{L}_{task}$, 2) a loss function associated with the domain classifier, which could be the binary cross-entropy function (or log loss, etc...) and denoted as $\mathcal{L}_{domain}$, and 3) a loss function associated with the reconstruction of the input data, denoted as $\mathcal{L}_{recon}$, and could be the mean-squared error (or hinge loss, etc...). The model tries to minimize the sum of the 3 loss functions, *i.e.* it wants to find the parameters θ^* such that:

$$\theta^* = \arg\min_{\theta} \; \mathcal{L}_{task} + \lambda \cdot \mathcal{L}_{domain} + \mu \cdot \mathcal{L}_{recon} \quad (1)$$

where λ and μ are real numbers in the range $[0, 1]$. Since the reconstruction error tends to be larger than the other 2 losses by orders of magnitude, its corresponding scalar μ tends to be small.

3.1 Label Predictor

Using the label predictor, the model predicts the labels of the input data. During training, since only the source domain data has labels, the input is sliced in a way that the N_s observations associated with the source domain are passed into the label predictor. The loss function $\mathcal{L}_{task}$ depends on the task at hand (Janocha and Czarnecki, 2017). For example, one could use the mean squared error for regression, or the binary cross entropy for classification. For our purpose, we use the binary cross entropy.

3.2 Domain Classifier

The model above should be robust towards shift in data distribution. Said differently, the model should be able to predict accurately the label of a given observation even when it comes from the target domain instead of the source domain. Mathematically, this is equivalent to minimizing the

error on label prediction and maximizing the error on domain classification. Ganin et al. (2016) showed that this can be done using a special layer they called Gradient Reversal Layer (GRL). The GRL does not affect the network during forward propagation, but it flips the sign of the gradients in backpropagation. The domain loss $\mathcal{L}_{domain}$ adopted by (Ganin et al., 2016) is the log-loss between the true domain and the predicted domain. Other binary loss functions are possible (Janocha and Czarnecki, 2017). In our approach, we use the binary cross entropy. The error of the domain classifier is scaled by λ.

3.3 Denoising Autoencoder

The noised version of X, denoted $\tilde{X}$, is obtained from X by using a masking noise, *i.e.* some elements of X are set to 0 with probability p (Glorot et al., 2011). Then, $\tilde{X}$ is propagated through an encoder network $h(\cdot)$ (Baldi, 2012) to get $h(\tilde{X})$. The decoder network $r(\cdot)$ reconstructs the input data X from the encoder's output $h(\tilde{X})$. A possible loss function is the mean squared error

$$\mathcal{L}_{recon} = \|r\left(h(\tilde{X})\right) - X\|^2 \quad (2)$$

The error of the autoencoder is scaled by μ.

3.4 Proxy A-distance as a Generalization Metric

Ben-David et al. (2007) developed a distance metric called proxy A-distance. The lower the distance, the more similar the domains are. Intuitively, this would mean the source and target domains share more common features. Hence, machine learning models won't lose too much accuracy when trained over source domain and tested over the target domain.

Let $\mathcal{D}$ and $\mathcal{D}'$ be 2 probability distributions defined over a domain χ, and a hypothesis class $\mathcal{A}$. The A-distance of $\mathcal{D}$ and $\mathcal{D}'$ is defined as

$$d_{\mathcal{A}}(\mathcal{D}', \mathcal{D}') = 2 \sup_{A \in \mathcal{A}} |\Pr_{\mathcal{D}}[A] - \Pr_{\mathcal{D}'}[A]| \quad (3)$$

Intuitively, this is equivalent to finding the maximum $L1$ distance between the 2 probability distributions $\mathcal{D}$ and $\mathcal{D}'$. Since computing this metric is intractable, Ben-David et al. (2007) proposed a way to approximate it from finite samples as follows: a linear SVM is trained to discriminate between the 2 domains, then the error ϵ, called generalization error, is used to compute a proxy of the A-distance $\hat{d}_{\mathcal{A}} = 2(1-2\epsilon)$. This proxy A-distance (PAD) can then be used to represent the distance between the 2 domains.

4 Experiments and Results

To test the effectiveness of the proposed approach, we conduct a 5-point sentiment classification on ArSentD-LEV[1] (Baly et al., 2018), once using the country of origin of the tweet as domain, and once the category to which the tweet belongs. We then show the effect of the data size on the performance of the adaptation algorithms. We start by describing the available dataset, then we describe each experiment alongside its results and we include some insights.

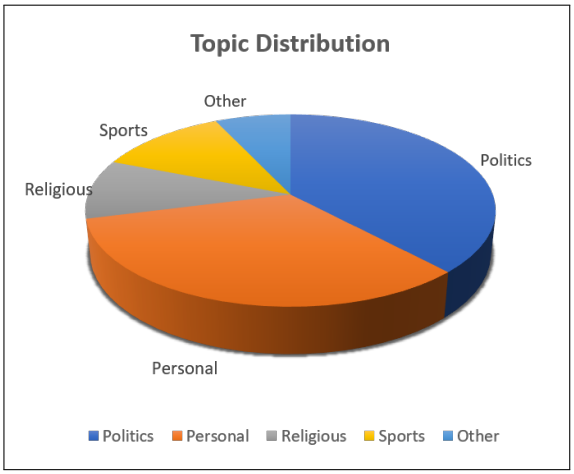

Figure 2: Topic Distribution of Tweets in ArSentD-LEV

4.1 Dataset Description and Experiment Setup

ArSentD-LEV is a multi-domain dataset containing almost 4,000 tweets collected equally from the 4 Levantine countries: Jordan, Lebanon, Palestine and Syria. For each tweet, the following labels are available: the country of origin, the sentiment conveyed by the tweet on 5-point scale (from very negative to very positive), the way of expressing the sentiment (explicit vs implicit) and the category to which the tweet belongs. The tweets were divided into 5 categories: politics, personal, sports, religious and other. The distribution of the tweets amongst these categories is shown in figure 2.

[1]The dataset is publicly available at http://oma-project.azurewebsites.net/ArSenL/ArSenTD_Lev_Intro

Following the approach used by (Chen et al., 2012; Ganin et al., 2016), we extract from the dataset the 5,000 most frequent unigrams and bigrams, as was adopted in (Ganin et al., 2016) for English. We then form, using these unigrams and bigrams, a bag-of-words matrix that will be used as input data for the learned models. Although many models represent text better (e.g. sequence models, tree models, etc...) we limit ourselves to a simpler model to show the improvement by the domain adaptation technique rather than by the text model.

The different experiments evaluated the performance of four models. A Linear SVM was used as a baseline and representative of feature based models. For the deep learning models, we consider a fully-connected neural network (Rumelhart et al., 1988) consisting of a hidden layer of 100 neurons and a label predictor of size 2. The setup of DANN is similar to that in (Ganin et al., 2016). The hidden layer is composed of 100 neurons, and the label predictor is of size 2. The domain classifier of DANN (of size 2) is preceded by a GRL. The proposed model is identical to the description in section 3. All neural networks were trained using ADAM optimizer (Kingma and Ba, 2014) using a learning rate of 10^{-3}.

Source	Target	*SVM*	*NN*	*DANN*	*Prop*
Jordan	**Lebanon**	**30**	27.5	29	**30**
	Palestine	33.5	33	34.5	**35**
	Syria	30.5	31.5	32	**33**
Lebanon	**Jordan**	**32**	28	29	**32**
	Palestine	**35**	25.5	31	**35**
	Syria	30.5	33	37	**37.5**
Palestine	**Jordan**	29.5	31	32	**32.5**
	Lebanon	**32**	29.5	31	31
	Syria	**37.5**	21.5	28.5	27.5
Syria	**Jordan**	**32.5**	32	30.5	32
	Lebanon	35	31.5	35	**35.5**
	Palestine	37	28	31.5	**37.5**

Table 1: Accuracies of linear SVM, NN, DANN and the proposed approach for Cross-Country adaptation on ArSentD-LEV. We can see that the proposed variant outperforms other models in almost all DA tasks.

4.2 Evaluation for Cross Country Adaptation

For this experiment, we evaluate the adaptation task between tweets from different countries. This means the source domain will consist of tweets from one of the 4 Levantine countries, and the target domain will consist of tweets coming from other countries. We thus have a total of 12 adaptation tasks. Baly et al. showed that Twitter is used for different purposes in different countries (Baly et al., 2017), which presents an additional challenge.

The result of the domain adaptation tasks are shown in Table 1. The proposed method outperformed all other models in most of the adaptation tasks. Although many real-life applications showed that traditional machine learning models are usually better when the available data is little (Cortes and Vapnik, 1995; Goodfellow et al., 2016), the proposed model was able to outperform the linear SVM in most of the tasks in our experiment. This means it was able to extract useful representation from the data. The model was also able to outperform DANN, which shows that the representational learning provides intrinsic representation of the data.

4.3 Evaluation for Cross Topic Adaptation

In this second experiment, we consider the task of adapting tweets from different topics. ArSentD-LEV (Baly et al., 2018) contains 5 classes for topic: politics, personal, religious, sports and other. This means we have a total of 20 tasks. The models evaluated are the linear SVM, DANN and the proposed model. The models' structure is identical to the one defined in section 4.2.

The results of the experiment are shown in Table 2. The behavior of the algorithms is significantly different in these categories. This is caused by the unbalanced data distribution amongst the different topics, as can be seen in Figure 2. We can see that whenever the data is very limited, the linear SVM outperforms the deep learning models. This is expected since neural networks cannot learn well the underlying representation when the data is scarce.

Looking at the radar plot in Figure 3, we can find the following interesting property. The higher the PAD distance between the source and target domains, the better the performance of the proposed model. This can be related to the fact that the proposed model tries to find a hidden representation that combines features from both source and target domains, *i.e.* decrease the distance between the 2 domains. Whenever the distance is low, the proposed model can not thus decrease it much further.

Source	Target	*SVM*	*DANN*	*Prop*
Politics	**Personal**	29.5	28.7	**33.3**
	Religious	20.5	20.3	**25.3**
	Sports	26.8	**35.1**	**35.1**
	Other	16.1	22.5	**24.2**
Personal	**Politics**	37.5	**41.7**	36.8
	Religious	19	22.8	**23.4**
	Sports	**34**	26.8	25.8
	Other	**40.3**	33.8	35.4
Religious	**Politics**	**16.8**	15.5	15.5
	Personal	**26.4**	24.1	26.1
	Sports	**28.8**	25.8	26.8
	Other	**48.4**	30.6	27.4
Sports	**Politics**	**41.4**	36.4	30.7
	Personal	**28.4**	25.3	24.5
	Religious	16.5	**20**	19
	Other	33.8	**35.5**	**35.5**
Other	**Politics**	20.5	**23.2**	**23.2**
	Personal	28.4	**30.3**	24.9
	Religious	**53.2**	41.8	43
	Sports	26.8	23.7	**27.8**

Table 2: Accuracies of linear SVM, DANN and the proposed approach for Cross-Topic on ArSentD-LEV. We can see that the SVM and the proposed variant are performing better than DANN, with SVM performing better when available data is little.

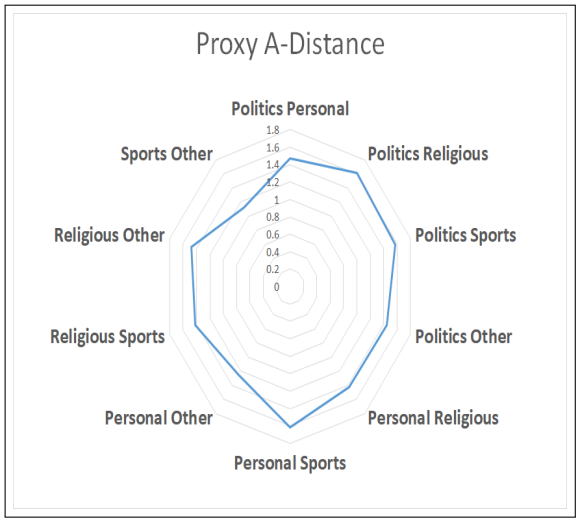

Figure 3: Proxy A-distance Between Different Domains. This radar plot shows the proxy A-distances between the different domains. The closer the vertex of a combination to the center, the closer the 2 domains.

4.4 Performance with Limited Data Size

To test the limitation of the proposed approach with data size, we consider the task where the source domain is "Politics" and the target domain is "Personal", since the available data is larger than the data available for other tasks. We then start by gradually increasing the size, and test the performance of the model with each dataset size. Looking at Figure 4, we can see that the performance of the proposed method is better than that of DANN at all sizes. This confirms our assumption that DANN with SDA learns a better representation through the incorporation of autoencoder. In contrast, DANN focuses on the discriminative task at hand, and thus fails to generalize. We also have a generally increasing trend which comes from the fact that more data is available, hence the models are able to learn better features.

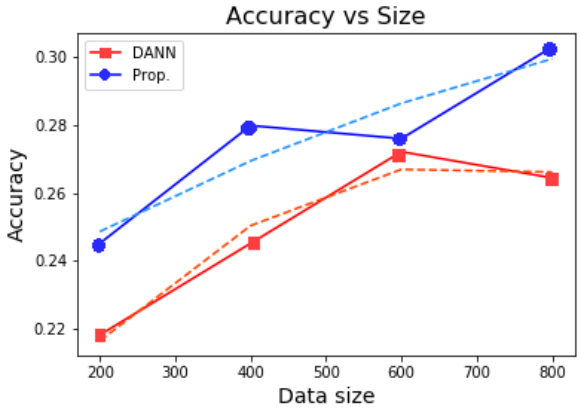

Figure 4: DANN and Proposed Method Performance vs Data size. We can see that the proposed variant outperforms DANN at all data sizes, and learns more with the increase in data size.

5 Conclusion

In this paper, we presented the first application of domain adaptation to the Arabic language. Although there exists work in English for domain adaptation, no work exists for Arabic. We considered in this paper the Domain Adversarial Neural Network (DANN) (Ganin et al., 2016) and proposed a variant that incorporates into DANN a stacked denoising autoencoder (SDA). The experiments and results provided several insights. We observed that integrating a reconstruction loss into DANN helped the model learn a better latent representation. This proved useful in all experiments, especially when the available data is little. These

observations are consistent with what has been observed in English. The success of domain adaptation suggests the possibility of usage of DA to bridge the gap between different dialects of the Arabic language. Future work includes testing DA techniques to more Arabic dialects, trying other domain adaptation algorithms in Arabic, developing new domain adaptation techniques, evaluating the DA tasks using better text representation (*e.g.* sequence models...) and integrating transfer learning techniques in the models (Ng et al., 2015).

References

Gilbert Badaro, Ramy Baly, Hazem Hajj, Wassim El-Hajj, Khaled Shaban, Nizar Habash, Ahmad 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 Transactions on Asian Language Information Processing*, 18.

Pierre Baldi. 2012. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, volume 27 of *Proceedings of Machine Learning Research*, pages 37–49, Bellevue, Washington, USA. PMLR.

Ramy Baly, Gilbert Badaro, Georges El-Khoury, Rawan Moukalled, Rita Aoun, Hazem Hajj, Wassim El-Hajj, Nizar Habash, and Khaled Shaban. 2017. A characterization study of Arabic twitter data with a benchmarking for state-of-the-art opinion mining models. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 110–118, Valencia, Spain. Association for Computational Linguistics.

Ramy Baly, Alaa Khaddaj, Hazem Hajj, Wassim El-Hajj, and Khaled Shaban. 2018. Arsentd-lev: A multi-topic corpus for target-based sentiment analysis in arabic levantine tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

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):151–175.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2007. Analysis of representations for domain adaptation. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 137–144. MIT Press.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447. 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. 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.

Minmin Chen, Zhixiang Xu, Kilian Q. Weinberger, and Fei Sha. 2012. Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 1627–1634, USA. Omnipress.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. In *Machine Learning*, pages 273–297.

Hal Daume III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

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.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 513–520, USA. Omnipress.

Boqing Gong, Kristen Grauman, and Fei Sha. 2013. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 222–230, Atlanta, Georgia, USA. PMLR.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2672–2680, Cambridge, MA, USA. MIT Press.

Katarzyna Janocha and Wojciech Czarnecki. 2017. On loss functions for deep neural networks in classification. 25.

Serena Jeblee, Weston Feely, Houda Bouamor, Alon Lavie, Nizar Habash, and Kemal Oflazer. 2014. Domain and dialect adaptation for machine translation into egyptian arabic. In *ANLP@EMNLP*.

Shafiq Joty, Preslav Nakov, Llus Mrquez, and Israa Jaradat. 2017. Cross-language learning with adversarial neural networks. pages 226–237.

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

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2018. Conditional adversarial domain adaptation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1640–1650. Curran Associates, Inc.

Will Monroe, Spence Green, and Christopher D. Manning. 2014. Word segmentation of informal arabic with domain adaptation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 206–211. Association for Computational Linguistics.

Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. 2015. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 443–449, New York, NY, USA. ACM.

Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 751–760, New York, NY, USA. ACM.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1988. Neurocomputing: Foundations of research. chapter Learning Representations by Back-propagating Errors, pages 696–699. MIT Press, Cambridge, MA, USA.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408.