Improving Cross-Lingual Sentiment Analysis via Conditional Language Adversarial Nets

Sai Hemanth Kandula
Tufts University
Sai_Hemanth.Kandula@tufts.edu

Bonan Min
Raytheon BBN Technologies
bonan.min@raytheon.com

Abstract

Sentiment analysis has come a long way for high-resource languages due to the availability of large annotated corpora. However, it still suffers from lack of training data for low-resource languages. To tackle this problem, we propose Conditional Language Adversarial Network (CLAN), an end-to-end neural architecture for cross-lingual sentiment analysis without cross-lingual supervision. CLAN differs from prior work in that it allows the adversarial training to be conditioned on both learned features and the sentiment prediction, to increase discriminativity for learned representation in the cross-lingual setting. Experimental results demonstrate that CLAN outperforms previous methods on the multilingual multi-domain Amazon review dataset. Our source code is released at https://github.com/hemanthkandula/clan.

1 Introduction

Recent success in sentiment analysis (Yang et al., 2019; Sun et al., 2019; Howard and Ruder, 2018; Brahma, 2018) is largely due to the availability of large-scale annotated datasets (Maas et al., 2011; Zhang et al., 2015; Rosenthal et al., 2017). However, such success can not be replicated to low-resource languages because of the lack of labeled data for training Machine Learning models.

As it is prohibitively expensive to obtain training data for all languages of interest, cross-lingual sentiment analysis (CLSA) (Barnes et al., 2018; Zhou et al., 2016b; Xu and Wan, 2017; Wan, 2009; Demirtas and Pechenizkiy, 2013; Xiao and Guo, 2012; Zhou et al., 2016a) offers the possibility of learning sentiment classification models for a target language using only annotated data from a different source language where large annotated data is available. These models often rely on bilingual lexicons, pre-trained cross-lingual word embeddings, or Machine Translation to bridge the gap between the source and target languages. CLIDSA/CLCDSA (Feng and Wan, 2019) is the first end-to-end CLSA model that does not require cross-lingual supervision which may not be available for low-resource languages.

In this paper, we propose Conditional Language Adversarial Network (CLAN) for end-to-end CLSA. Similar to prior work, CLAN performs CLSA without using any cross-lingual supervision. Differing from prior work, CLAN incorporates conditional language adversarial training to learn language invariant features by conditioning on both learned feature representations (or features for short) and sentiment predictions, therefore increases the features' discriminativity in the cross-lingual setting. Our contributions are three fold:

- We develop Conditional Language Adversarial Network (CLAN) which is designed to learn language invariant features that are also discriminative for sentiment classification.
- Experiments on the multilingual multi-domain Amazon review dataset (Prettenhofer and Stein, 2010) show that CLAN outperforms all previous methods for both in-domain and cross-domain CLSA tasks.
- t-SNE visualization of the held-out examples shows that the learned features align well across languages, indicating that CLAN is able to learn language invariant features.

2 Related Work

Cross-lingual sentiment analysis (CLSA): Several CLSA methods (Wan, 2009; Demirtas and Pechenizkiy, 2013; Xiao and Guo, 2012; Zhou et al., 2016a; Wan, 2009; Xiao and Guo, 2012) rely on Machine Translation (MT) for providing supervision across languages. MT, often trained from parallel corpora, may not be available for low-resource languages. Other CLSA methods (Barnes et al., 2018; Zhou et al., 2016b; Xu and Wan, 2017)

Proceedings of the Third Workshop on Computational Typology and Multilingual NLP, pages 32–37

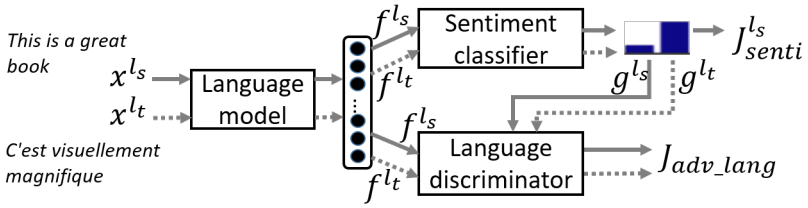

Figure 1: CLAN architecture. We illustrate with a source language l_s =English (solid line) and target language l_t =French (dotted line). x^{l_s}, x^{l_t} are sentences in l_s and l_t, f^{l_s}, f^{l_t} are features extracted by the language model for x^{l_s} and x^{l_t}, and g^{l_s}, g^{l_t} are the sentiment predictions for x^{l_s} and x^{l_t}, respectively. The sentiment classification loss $\mathcal{J}^{l_s}_{senti}$ is only trained on x^{l_s} for which the sentiment label is available, while the language discriminator is trained from both x^{l_s} and x^{l_t}.

uses bilingual lexicons or cross-lingual word embeddings (CLWE) to project words with similar meanings from different languages into nearby spaces, to enable training cross-lingual sentiment classifiers. CLWE often depends on a bilingual lexicon (Barnes et al., 2018) or parallel or comparable corpora (Mogadala and Rettinger, 2016; Vulić and Moens, 2016). Recently, CLWE methods (Lample and Conneau, 2019; Conneau et al., 2019) that rely on no parallel resources are proposed, but they require very large monolingual corpora to train. The work that is most related to ours is (Feng and Wan, 2019), which does not rely on cross-lingual resources. Different from the language adversarial network used in (Feng and Wan, 2019), our work performs cross-lingual sentiment analysis using conditional language adversarial training, which allows the language invariant features to be specialized for sentiment class predictions.

Adversarial training for domain adaptation Our approach draws inspiration from Domain-Adversarial Training of Neural Networks (Ganin et al., 2016) and Conditional Adversarial Domain Adaptation (CDAN) (Long et al., 2018). DANN (Ganin et al., 2016) trains a feature generator to minimize the classification loss, and a domain discriminator to distinguish the domain where the input instances come from. It attempts to learn domain invariant features that deceive the domain discriminator while learning to predict the correct sentiment labels.CDAN (Long et al., 2018) additionally makes the discriminator conditioned on both extracted features and class predictions to improve discriminativity.

3 Conditional Language Adversarial Networks for Sentiment Analysis

Figure 1 shows the architecture of CLAN. It has three components: a multilingual language model (LM) that extracts features from the input sentences, a sentiment classifier built atop of the features extracted by the LM, and a conditional language adversarial trainer to force the features to be language invariant. All three components are jointly optimized in a single end-to-end neural architecture, allowing CLAN to learn cross-lingual features and to capture multiplicative interactions between the features and sentiment predictions. The resulting cross-lingual features are specialized for each sentiment class.

CLAN aims at solving the cross-lingual multi-domain sentiment analysis task. Formally, given a set of domains $\mathcal{D}$ and a set of languages $\mathcal{L}$, CLAN consists of the following components:

- Sentiment classifier: train on (l_s, d_s) (sentiment labels are available) and test on (l_t, d_t) (no sentiment labels), in which $l_s, l_t \in \mathcal{L}, l_s \neq l_t$ and $d_s, d_t \in \mathcal{D}$. CLAN works for both variants of the CLSA problem: **in-domain CLSA** where $d_s = d_t$, and **cross-domain CLSA** where $d_s \neq d_t$.
- Language model: train on (l, d) in which $l \in \mathcal{L}, d \in \mathcal{D}$.
- Language discriminator: train on (l, d) in which $l \in \mathcal{L}$ and $d \in \mathcal{D}$. The language IDs are known.

Language Model (LM): For a sentence x, we compute the probability of seeing a word w_k given the previous words: $p(x) = \prod_{k=1}^{|x|} P(w_k|w_1, ..., w_{k-1})$: we first pass the input words through the embedding layer of language l parameterized by θ^l_{emb}. The embedding for word w_k is $\vec{w_k}$. We then pass the word embeddings to two LSTM layer parameterized by θ_1 and θ_2, that are shared across all languages and all domains, to generate hidden states $(z_1, z_2, ..., z_x)$ that can be considered as features for CLSA: $h_k = \text{LSTM}(h_{k-1}, \vec{w_k}; \theta_1)$, and $z_k = \text{LSTM}(z_{k-1}, h_k; \theta_2)$. We then use a linear decoding layer parameterized by θ^l_{dec} with a softmax for next word prediction. To summarize, the LM objective for l is:

$$\mathcal{J}^l_{lm}(\theta^l_{emb}, \theta_1, \theta_2, \theta^l_{dec}) = \mathbb{E}_{x\sim\mathcal{L}^l}[-\frac{1}{|x|}\sum_{k=1}^{|x|} \log p(w_k|w1, ..., w_{k-1})]$$

where $x \sim \mathcal{L}^l$ indicates that x is sampled from text in language l.

Sentiment Classifier We use a linear classifier that takes the average final hidden states $\frac{1}{|x|}\sum_{k=1}^{|x|} z_k$ as input features, and then a softmax to output sentiment labels. The objective is:

$$\mathcal{J}^l_{senti}(\theta^l_{emb}, \theta_1, \theta_2, \theta^l_{senti}) = \mathbb{E}_{(x,y)\sim\mathcal{C}_{senti}}[-\log p(y|x)]$$

where $(x, y) \sim \mathcal{C}^l_{senti}$ indicates that the sentence x and its label y are sampled from the labeled examples in language l, and θ^l_{senti} denotes the parameters of the linear sentiment classifier.

Conditional Language Adversarial Training To force the features to be language invariant, we adopted conditional adversarial training (Long et al., 2018): a language discriminator is trained to predict language ID given the features by minimizing the cross-entropy loss, while the LM is trained to fool the discriminator by maximizing the loss:

$$\mathcal{J}^l_{adv\_lang}(\theta_{emb}, \theta_1, \theta_2, \theta_{dis\_lang}) = \mathbb{E}_{(x,l)}[-\log p(l|f(x) \otimes g(x))]$$

where $f(x)$, $g(x)$ and $l \in L$ are features extracted by the LM for input sentence x, its sentiment prediction and its language ID respectively, $\theta_{emb} = \theta^1_{emb} \oplus \theta^2_{emb} \oplus ... \oplus \theta^{|\mathcal{L}|}_{emb}$ denotes the parameters of all embedding layers and θ_{dis_lang} denotes the parameters of the language discriminator. We use multilinear conditioning (Long et al., 2018) by conditioning l on the cross-covariance $f(x) \otimes g(x)$.

A key innovation is the conditional language adversarial training: the multilinear conditioning enables manipulation of the multiplicative interactions across features and class predictions. Such interactions capture the cross-covariance between the language invariant features and classifier predictions to improve discriminability.

The Full Model Putting all components together, the final objective function is the following:

$$\mathcal{J}(\theta_{emb}, \theta_{lstm}, \theta_{dec}, \theta_{senti}, \theta_{dis\_lang}) = \sum_{(l,d)} \mathcal{J}^l_{lm} + \alpha\mathcal{J}^l_{senti} - \beta\mathcal{J}^l_{adv\_lang}$$

where $\theta_{lstm} = \theta_1 \oplus \theta_2$ denotes parameters of the LSTM layers, $\theta_{dec} = \theta^1_{dec} \oplus \theta^2_{dec} \oplus ... \oplus \theta^{|\mathcal{L}|}_{dec}$ denotes the paramters of all decoding layers, α and β are hyperpameters controlling the relative importance of the sentiment classification and the language adversarial training objectives. Parameters θ_{dis_lang} is trained to maximize the full objective function while the others are trained to minimize it:

$$\hat{\theta}_{dis\_lang} = \arg\max_{\theta_{dis\_lang}} \mathcal{J}$$

$$(\hat{\theta}_{emb}, \hat{\theta}_{lstm}, \hat{\theta}_{dec}, \hat{\theta}_{senti}) = \arg\min_{\theta_{emb}, \theta_{lstm}, \theta_{dec}, \theta_{senti}} \mathcal{J}$$

4 Experiments

Datasets: We evaluate CLAN on the Websis-CLS-10 dataset (Prettenhofer and Stein, 2010) which consists of Amazon product reviews from 4 languages and 3 domains. Following prior work, we use English as the source language and other languages as the target languages. For each language-domain pair there are 2,000 training documents, 2,000 test documents, and 9,000-50,000 unlabeled documents depending on the language-domain pair (details are in Prettenhofer and Stein, 2010).

Implementation details: The models are implemented in PyTorch(Paszke et al., 2019). All models are trained on four NVIDIA 1080ti GPUs. We tokenized text using NLTK (Loper and Bird, 2002). For each language, we kept the most frequent 15000 words in the vocabulary since a bigger vocabulary leads to under-fitting and much longer training time. We set the word embedding size to 600 for the language model, and use 300 neurons for the hidden layer in the sentiment classifier. We set $\alpha = 0.02$ and $\beta = 0.1$ for all experiments. All weights of CLAN were trained end-to-end using Adam optimizer with a learning rate of 0.03. We train the models with a maximum of 50,000 iterations with early stopping (typically stops at 3,000-4,000 iterations) to avoid over-fitting.

Experiment results: We follow the experiment setting described in (Feng and Wan, 2019). Table 1a and 1b show the accuracy of CLAN comparing to prior methods for the in-domain CLSA and cross-domain CLSA tasks, respectively. We compare CLAN to the following methods: CL-SCL, BiDRL, UMM, CLDFA, CNN-BE (Ziser and Reichart, 2018), PBLM-BE (Ziser and Reichart, 2018), A-SCL (Ziser and Reichart, 2018) are methods that require cross-lingual supervision

	English-German				English-French				English-Japanese			
	B	**D**	**M**	**AVG**	**B**	**D**	**M**	**AVG**	**B**	**D**	**M**	**AVG**
CL-SCL (Prettenhofer and Stein, 2010)	79.5	76.9	77.7	78.0	78.4	78.8	77.9	78.3	73.0	71.0	75.1	73.0
BiDRL (Zhou et al., 2016a)	84.1	84.0	84.6	84.2	84.3	83.6	82.5	83.4	73.1	76.7	78.7	76.1
UMM (Xu and Wan, 2017)	81.6	81.2	81.3	81.3	80.2	80.2	79.4	79.9	71.2	72.5	75.3	73.0
CLDFA (Xu and Yang, 2017)	83.9	83.1	79.0	82.0	83.3	82.5	83.3	83.0	77.3	80.5	76.4	78.0
MAN-MoE (Chen et al., 2019)	82.4	78.8	77.1	79.4	81.1	84.2	80.9	82.0	62.7	69.1	72.6	68.1
MWE (Conneau et al., 2017)	76.1	76.8	74.7	75.8	76.3	78.7	71.6	75.5	-	-	-	-
CLIDSA (Feng and Wan, 2019)	86.6	**84.6**	85.0	85.4	87.2	87.9	87.1	87.4	79.3	81.9	84.0	81.7
CLAN	**88.2**	84.5	**86.3**	**86.3**	**88.6**	**88.7**	**87.7**	**88.3**	**82.0**	**84.1**	**85.1**	**83.7**

(a) Accuracy for in-domain CLSA.

	English-German							English-French						
$S \rightarrow T$	**D→B**	**M→B**	**B→D**	**M→D**	**B→M**	**D→M**	**AVG**	**D→B**	**M→B**	**B→D**	**M→D**	**B→M**	**D→M**	**AVG**
CNN-BE	62.8	63.8	65.3	68.7	71.6	72.0	67.3	69.5	59.7	63.7	65.7	65.9	67.0	65.2
DCI	67.1	60.6	66.9	66.7	68.9	68.2	66.4	71.2	65.4	69.1	67.5	66.7	71.4	68.6
CL-SCL	65.9	62.5	65.1	65.2	71.2	69.8	66.7	70.3	63.8	68.8	66.8	66.0	70.1	67.6
A-SCL	67.9	63.7	68.7	63.8	69.0	70.1	67.2	68.6	66.1	69.2	69.4	66.7	68.1	68.0
A-S-SR	68.3	62.5	69.4	69.9	70.2	69	67.4	69.3	68.9	70.9	70.7	67	71.4	69.7
PBLM+BE	78.7	78.6	80.6	79.2	81.7	78.5	79.5	81.1	74.7	76.3	75.0	75.1	76.8	76.5
CLCDSA	85.4	81.7	79.3	81.0	**83.4**	81.7	82.0	86.2	81.8	84.3	82.8	83.7	85.0	83.9
CLAN	**86.9**	**85.1**	**82.4**	**81.6**	83	**83.8**	**83.8**	**87.3**	**85.5**	**85.3**	**83.9**	**85.5**	**85.7**	**85.5**

(b) Accuracy for cross-domain CLSA. Six domain pairs were generated for each language pair. S and T refers to the source and target domains, respectively.

Table 1: Accuracy of CLSA methods on Websis-CLS-10. Top scores are shown in bold. D, M, B refers to DVD, music, and books, respectively. AVG refers to the average of scores per each language pair.

such as bilingual lexicons or Machine Translation. MAN-MoE and MWE use MUSE (Conneau et al., 2017) to generate cross-lingual word embeddings. CLIDSA/CLCDSA (Feng and Wan, 2019) uses language adversarial training. We refer readers to the corresponding papers for details of each model.

As shown in Table 1a and 1b, CLAN outperforms all prior methods in 11 out of 12 settings for cross-domain CLSA, and outperforms all prior methods in 8 out of 9 settings for in-domain CLSA. On average, CLAN achieves state-of-the-art performance on all language pairs for both in-domain and cross-domain CLSA tasks.

Analysis of results: To understand what features CLAN learned to enable CLSA, we probed CLAN by visualizing the distribution of features extracted from held-out examples from the language model through t-SNE (Maaten and Hinton, 2008). The plots are in Figure 2. The t-SNE plots show that the feature distributions for sentences in the source and target languages align well, indicating that CLAN is able to learn language-invariant features. To further look into what CLAN learns, we manually inspected 50 examples where CLAN classified correctly but the prior models failed: for example, in the books domain in German, CLAN classified "*unterhaltsam und etwas lustig*" ("*entertaining and a little funny*") correctly as positive, also classified the following text correctly as positive: "*ein buch dass mich gefesselt hat...Dieses Buch ist absolut nichts für schwache Nerven oder Moralisten*" ("*a book that captivated me...this book is absolutely not for the faint of heart or moralists!*"). This indicates that CLAN is able to learn better lexical, syntactic and semantic features.

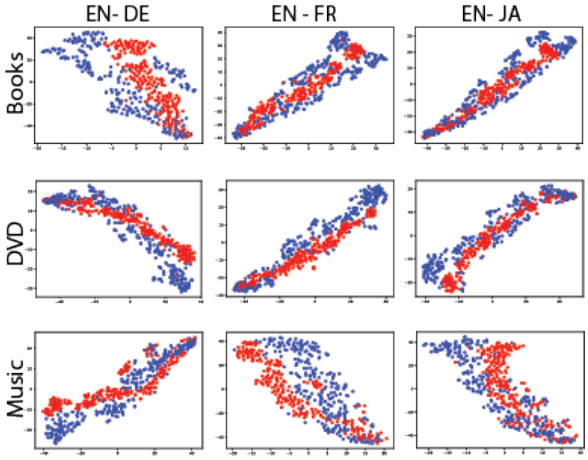

Figure 2: t-SNE plots of the distributions of features extracted from CLAN's language model, trained via the in-domain CLSA task. Red and blue dots represent features extracted from the source and target language held-out sentences, respectively. EN, DE, FR, JA refers to English, German, French and Japanese respectively.

5 Conclusion

We present Conditional Language Adversarial Networks for cross-lingual sentiment analysis, and show that it achieves state-of-the-art performance.

Acknowledgements

This work was supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2019-19051600006 under the BETTER program. The views, opinions, and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views or policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2483–2493, Melbourne, Australia. Association for Computational Linguistics.

Siddhartha Brahma. 2018. Improved sentence modeling using suffix bidirectional lstm. *arXiv preprint arXiv:1805.07340*.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Erkin Demirtas and Mykola Pechenizkiy. 2013. Cross-lingual polarity detection with machine translation. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, pages 1–8.

Yanlin Feng and Xiaojun Wan. 2019. Towards a unified end-to-end approach for fully unsupervised cross-lingual sentiment analysis. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 1035–1044, Hong Kong, China. Association for Computational Linguistics.

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.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

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.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, page 63–70, USA. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. 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.

Aditya Mogadala and Achim Rettinger. 2016. Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 692–702, San Diego, California. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1127, Uppsala, Sweden. Association for Computational Linguistics.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

Ivan Vulić and Marie-Francine Moens. 2016. Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, 55:953–994.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 235–243, Suntec, Singapore. Association for Computational Linguistics.

Min Xiao and Yuhong Guo. 2012. Multi-view AdaBoost for multilingual subjectivity analysis. In *Proceedings of COLING 2012*, pages 2851–2866, Mumbai, India. The COLING 2012 Organizing Committee.

Kui Xu and Xiaojun Wan. 2017. Towards a universal sentiment classifier in multiple languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 511–520, Copenhagen, Denmark. Association for Computational Linguistics.

Ruochen Xu and Yiming Yang. 2017. Cross-lingual distillation for text classification. *arXiv preprint arXiv:1705.02073*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016a. Attention-based LSTM network for cross-lingual sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 247–256, Austin, Texas. Association for Computational Linguistics.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016b. Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1412, Berlin, Germany. Association for Computational Linguistics.

Yftah Ziser and Roi Reichart. 2018. Deep pivot-based modeling for cross-language cross-domain transfer with minimal guidance. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 238–249, Brussels, Belgium. Association for Computational Linguistics.