FIT BUT at SemEval-2023 Task 12: Sentiment Without Borders - Multilingual Domain Adaptation for Low-Resource Sentiment Classification

Maksim Aparovich and **Santosh Kesiraju** and **Aneta Dufkova** and **Pavel Smrz**
Brno University of Technology, Faculty of Information Technology
612 66 Brno, Czech Republic
{iaparovich,kesiraju,smrz}@fit.vutbr.cz
xdufko02@stud.fit.vutbr.cz

Abstract

This paper presents our proposed method for SemEval-2023 Task 12, which focuses on sentiment analysis for low-resource African languages. Our method utilizes a language-centric domain adaptation approach which is based on adversarial training, where a *small* version of Afro-XLM-Roberta serves as a *generator* model and a feed-forward network as a *discriminator*. We participated in all three subtasks: monolingual (12 tracks), multilingual (1 track), and zero-shot (2 tracks). Our results show an improvement in weighted F1 for 13 out of 15 tracks with a maximum increase of 4.3 points for Moroccan Arabic compared to the baseline. We observed that using language family-based labels along with sequence-level input representations for the discriminator model improves the quality of the cross-lingual sentiment analysis for the languages unseen during the training. Additionally, our experimental results suggest that training the system on languages that are close in a language families tree enhances the quality of sentiment analysis for low-resource languages. Lastly, the computational complexity of the prediction step was kept at the same level which makes the approach to be interesting from a practical perspective. The code of the approach can be found in our repository [1].

1 Introduction

The goal of sentiment analysis is to determine the attitude or emotion expressed by the writer or speaker. The objective of sentiment analysis is to extract subjective information from text data, such as opinions, emotions, and intentions (Mukherjee and Bhattacharyya, 2013). Sentiment classification is a subtask of sentiment analysis that involves classifying the sentiment of a given text as either positive, negative, or neutral. The analysis of sentiment can be on document, sentence, phrase, or token level.

[1] https://github.com/KNOT-FIT-BUT/sentiment-without-borders

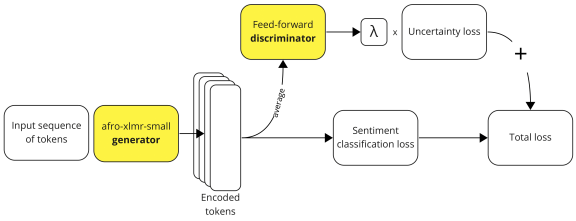

Figure 1: System architecture used in the paper.

The sentiment analysis of low-resource languages is challenging due to the limited availability of annotated data and the scarcity of language-specific resources. SemEval-2023 Task 12 (Muhammad et al., 2023b) and the Aftrisenti dataset (Muhammad et al., 2023a) alleviates this limitation for African languages.

The problem of dataset shift or domain shift can be mitigated by domain adaptation, which is a technique that deals with the fact that datasets may contain samples from different distributions. This issue can arise when certain domains have limited labeling or data, which results in a model failing to generalize well on those domains. Moreover, domain adaptation can improve out-of-distribution generalization and enable targeting of unknown domains (Volpi et al., 2018).

This paper describes an approach that is based on a language-centric domain adaptation with a small version of Afro-XLM-Roberta (Alabi et al., 2022) as a generator model and a feed-forward network as a discriminator. Our results show an improvement in weighted F1 for 13 out of 15 tracks with a maximum increase of 4.3 points for Moroccan Arabic compared to the baseline. In addition, it was observed that the quality of cross-lingual setup improves for unseen languages when language family-based labels are used in adversarial training. We have experimented with token and sequence-level discriminator inputs and found the latter one to be more favorable. Moreover, based on our experimental results, it can be suggested that training

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1518–1524

the model on languages that are closely related in the language family tree can enhance the quality of sentiment analysis for low-resource languages using language-centric domain adaptation.

2 Task Description

The goal of the SemEval-2023 Task 12: Sentiment Analysis for African Languages is to enable sentiment analysis research in African languages. The shared task is based on the Afrisenti dataset (Muhammad et al., 2023a) that consists of tweets in 14 African languages (Hausa, Yoruba, Igbo, Nigerian Pidgin, Amharic, Algerian Arabic dialect, Moroccan Arabic/Darija, Swahili, Kinyarwanda, Twi, Mozambique Portuguese, Xitsonga, Tigrinya, Oromo) from 4 branches of language families (Afro-Asiatic, English Creole, Indo-European, Niger-Congo). Each tweet has a positive, neutral, or negative label. The SemEval-2023 Task 12 has 3 subtasks:

1. Subtask A: Monolingual Sentiment Classification (all languages except Tigrinya and Oromo)
2. Subtask B: Multilingual Sentiment Classification (all languages except Tigrinya and Oromo)
3. Subtask C: Zero-Shot Sentiment Classification (Tigrinya and Oromo).

The submissions are ranked by weighted F1 score. Our team participated in all three subtasks with a single model trained on the Subtask B data.

3 Related Work

Previous works in sentiment analysis, including multilingual problems, relied on lexicon-based approaches (Nielsen, 2011), classical ML models with hand-crafted features (Dashtipour et al., 2016), or elder NN architectures (Kim, 2014). Recent approaches include the usage of pre-trained Transformer-based LMs (Sun et al., 2019; Yang et al., 2019).

Previous studies on domain adaptation for sentiment analysis mitigate the problem of domain shift aiming at building a model which is generalized to multiple domains (Ruder et al., 2017; Toledo-Ronen et al., 2022; Ganin et al., 2016; Guo et al., 2020).

In our work, we leverage a Transformer-based model pre-trained on African languages and adopt the same perspective in terms of generalization but from a different angle, treating languages and language families as domains. A similar usage of domain adaptation was proposed in Lample et al. (2018) for a two-language machine translation problem. In our study, we extend this language-centric approach to a multilingual setup and experiment with different options for discriminator input and domain labels.

4 System Overview

In this section, we first describe the main model which is also called the generator model. Next, we discuss variations of domain adaptation applications and options for discriminator input. A high-level schema of the final system used for submission is provided in Figure 1.

4.1 Generator Model

We utilized the Afro-XLM-Roberta (Alabi et al., 2022) as the generator model. It was created by reducing the vocabulary of the base version of XLM-Roberta (Conneau et al., 2020) from 250K to 70k tokens. This was followed by the multilingual adaptive finetuning adaptation process. The model has 12 transformer layers with a hidden size of 768. It was trained on 17 African languages. However, 6 languages and dialects from the Afrisenti dataset were not seen during the training phase: Algerian Arabic, Moroccan Arabic, Twi, Mozambican Portuguese, Xitsonga (Subtasks A and B), and Tigrinya (Subtask C). In order to experiment faster and obtain ablation study results for all the modifications we considered, a small version of the model was used.

4.2 Domain Adaptation

There are different ways to categorize domain adaptation, such as supervised (source and target labels are available) and unsupervised (unlabeled target data), and model-centric, data-centric, and hybrid approaches (Ramponi and Plank, 2020).

Our work focuses on model-centric domain adaptation with adversarial training, where we modified the loss function of the generator model by adding an adversarial component. Languages and language families were concerned as domains for adaptation and were used as target labels by the discriminator model.

4.2.1 Adversarial Loss

At a high level, each training step consists of two phases. In the first phase, the discriminator model learns to predict the correct domain class based on the latent text representations produced by the generator model. In the second phase, the generator model produces latent text representations which are used to predict sentiment class and to fool the discriminator at the same time. The more discriminator is uncertain in its predictions, the better.

The inspiration to use adversarial training came from Lample et al. (2018). The goal of the approach is to restrict the generator to output text representations from the same feature space for different domains. Then, the classification head responsible for sentiment classification would be able to operate with those representations regardless of the domain. As a domain, we considered two options: language and language family. In addition, a language script could be used as a domain, but we did not experiment in that direction.

Discriminator classifies a single vector $x, x \in \mathbb{R}^{gen.\ hidden\ size}$ at a time, which can be either (i) one of the latent text representations output by the generator or (ii) an average of the output sequence of latent text representations. Then, it outputs domain class probabilities $p(l|x)$, where $l, l \in [0, \ldots, L]$ is a domain class. The discriminator is trained to predict the correct domain class l by minimizing the cross-entropy loss:

$$\mathcal{L}_{disc} = -\frac{1}{N}\sum_{i=1}^{N} \log p(l_{true}|x_i), \qquad (1)$$

where N is a length of an input sequence.

Then, the generator is trained to predict sentiment and to fool the discriminator at the same time. Lample et al. (2018) approach uses a binary classification problem setup and forces the discriminator to predict an opposite label. The same setup can not be directly applied to multi-class classification. Intuitively, the goal of multi-class adversarial training is to make the discriminator to be equally uncertain in all the classes. Therefore, we consider the discriminator as fooled if produced class probabilities are close to $\frac{1}{L}$ which is used as target probabilities. The adversarial loss would have the following form:

$$\mathcal{L}_{adv} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=i}^{L} \frac{1}{L} \log p(l_j|x_i) \qquad (2)$$

The objective function of the main model is:

$$\mathcal{L}_{gen} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{adv}, \qquad (3)$$

where $\mathcal{L}_{cls}$ is a cross-entropy loss of sentiment classification and λ is an adversarial loss scaling factor.

4.2.2 Discriminator Input

In our paper, we explored the effects of applying domain adaptation at different levels of discriminator input, specifically on the token and sequence levels. Intuitively, our goal is for the generator to produce text representations that are language-agnostic, meaning they are consistent across different input languages. Domain adaptation on the token level should lead to better alignment in the generator latent text representations space. Conversely, applying domain adaptation at the sequence level may indirectly result in an improved generalization in a space of semantic meaning across languages. To investigate these hypotheses, we experimented with both approaches. In order to represent sequence-level we use a vector obtained as an average of token-level latent representations produced by the generator.

Parameter	Options	Best option
$batch\_size$	16, 32, 64, 128	32
$learning\_rate_{gen}$	1e-5, 2e-5, 3e-5	3e-5
$linear\_decay_{gen}$	True, False	True
$weight\_decay_{gen}$	1e-4, 1e-2, 1e-1, 0	0
$hidden\_size_{disc}$	192, 384, 768, 1536	768
$learning\_rate_{disc}$	1e-3, 3e-3, 1e-4, 3e-4	3e-4
$linear\_decay_{disc}$	True, False	True

Table 1: The table provides an overview of the hyperparameters used during the optimization process and highlights the best-performing values that were identified through fine-tuning. Hyperparameters were fine-tuned separately for the generator model (gen) and the discriminator (disc).

5 Experimental Setup

Our experiments focused on Subtasks B and C and used the available multilingual training data from Subtask B. The training, validation, and test datasets were provided by the task organizers and were used as is, without any additional preprocessing. Texts were tokenized by Afro-XLM-Roberta pre-trained SentencePiece tokenizer (Kudo and Richardson, 2018)

First, we performed hyperparameter tuning for the generator and discriminator models separately. We tuned the hyperparameters for the generator model using a randomly sampled 25% of the training data while keeping the validation dataset as is. We then repeated the same process for the discriminator model, using the same set of optimized hyperparameters for both models separately. A summary of the hyperparameters space and best-performing set of parameters can be found in Table 1.

We also prepared a mapping of languages to language families, using an ad-hoc approach. To group languages by family, we went up the language families tree until the Kth level and formed a group of languages that shared a common family. The value of K was empirically set to 5. It is possible that a group can have only one language.

While experimenting with the adversarial training setup, we investigated the effects of domain adaptation on the token and sequence levels, using either language or language family as labels for the discriminator, and varying the discriminator loss scaling factors. We used the same stop criterion for all experiments: if there were no improvement greater than 2 points in weighted F1 during 5 subsequent steps, training was stopped.

In our implementation, we utilized a pre-trained Afro-XLM-Roberta model and tokenizer from the Huggingface (Wolf et al., 2020) repository. Our training pipeline was developed using the PyTorch-lightning framework with PyTorch (Paszke et al., 2019) as a backend.

6 Results and Ablation Study

Results were submitted for the model trained with a set of hyperparameters that differ from the presented experimental results. During the pre-evaluation period, we experimented with generator warmup steps which resulted in better results for Subtask A, almost the same performance for Subtask B, and worse quality for Subtask C. Taking into account our focus on Subtasks B and C we decided not to use warmup steps in our post-evaluation experiments. As a result, the baseline and experiments are based on the same set of hyperparameters, and the submission model was trained on a different set. Results for submission, baseline, and experiments can be found in Table 2.

The baseline and experimental models had a batch size of 32, a generator learning rate of 3e-5 with linear decay, weight decay of 0, and maximum training steps of 150000.

The submitted model had a generator learning rate of 1e-4 with 2000 warmup steps and linear decay. The adversarial loss scaling factor λ was set to 1. Other parameters were preserved the same.

The discriminator had the same hyperparameters in both options and had a hidden layer size of 768 and a learning rate of 3e-4 with linear decay.

6.1 Adversarial Training

Domain adaptation setup beats baseline in all cases except Amharic (am) and Oromo (or) languages. The scores reported in Table 2 for each option represent the mean and standard deviation of three separate runs.

In Subtask C, zero-shot classification, Oromo does not outperform the baseline in contrast to Tigrinya. Interestingly, the former language was seen during the pre-training phase, while the latter one was not. Training data contains the Amharic language which is relatively close to Tigrinya in the language families tree – both are Ethiopic languages from the Afro-Asiatic branch. We hypothesize that domain adaptation in cooperation with having a close language helps in a zero-shot task setup and not having such language leads to faster degradation of performance.

6.2 Adversarial Loss Scaling

Our experimental results suggest that a discriminator loss scale within the [0, 1] range yields the best performance. Conversely, a larger value of 10 often leads to a degradation in quality across most cases.

6.3 Discriminator Labels

We found that using language-level labels outperforms the approach based on language families in the case of supervised learning. However, language families-based labels surpass language-based labels in cross-lingual setup for unseen languages from Subtask C (Table 3). We should note that our

Language	am (a)	dz (a)	ha (a)	ig (a)	kr (a)	ma (a)	pcm (a)
submission	65.1	62.2	72.8	75.6	**65.4**	51.2	64.9
baseline	$\mathbf{70.3 \pm 0.1}$	62.9 ± 0.9	72.1 ± 1.5	74.4 ± 1.0	65.1 ± 2.1	50.4 ± 0.2	65.7 ± 0.6
$DA_{\lambda=0.10}$	64.9 ± 4.3	62.3 ± 1.0	72.9 ± 2.2	73.1 ± 1.9	64.5 ± 2.3	53.9 ± 2.2	65.9 ± 0.3
$DA_{\lambda=0.25}$	61.4 ± 5.6	63.1 ± 0.8	73.8 ± 2.0	74.2 ± 1.6	63.1 ± 2.0	52.5 ± 3.1	$\mathbf{66.2 \pm 1.2}$
$DA_{\lambda=0.50}$	61.2 ± 7.0	61.3 ± 1.2	73.4 ± 2.4	$\mathbf{74.9 \pm 2.0}$	64.0 ± 1.3	$\mathbf{54.7 \pm 1.4}$	65.7 ± 1.1
$DA_{\lambda=0.75}$	63.7 ± 1.7	$\mathbf{63.2 \pm 0.3}$	74.3 ± 0.5	74.3 ± 0.4	64.7 ± 1.5	52.8 ± 1.1	65.8 ± 0.5
$DA_{\lambda=1.00}$	66.0 ± 5.2	62.0 ± 0.6	$\mathbf{75.1 \pm 0.3}$	$\mathbf{74.9 \pm 1.4}$	$\mathbf{65.4 \pm 0.5}$	50.5 ± 0.8	65.9 ± 1.1
$DA_{\lambda=10.00}$	59.8 ± 1.0	62.8 ± 1.0	73.0 ± 1.5	73.0 ± 1.5	64.2 ± 2.3	54.4 ± 2.6	65.4 ± 0.7

Language	pt (a)	sw (a)	ts (a)	twi (a)	yo (a)	mult. (b)	or (c)	tg (c)
submission	**63.6**	58.4	47.4	**63.4**	68.6	**66.5**	38.1	50.9
baseline	60.9 ± 0.4	57.4 ± 3.4	49.7 ± 1.3	61.6 ± 2.2	67.4 ± 2.0	66.1 ± 0.7	$\mathbf{39.0 \pm 1.3}$	51.0 ± 2.4
$DA_{\lambda=0.10}$	59.3 ± 2.6	55.4 ± 1.0	$\mathbf{53.2 \pm 3.2}$	61.6 ± 1.0	66.9 ± 2.6	65.6 ± 0.9	37.6 ± 2.3	53.7 ± 0.8
$DA_{\lambda=0.25}$	$\mathbf{62.7 \pm 0.9}$	56.7 ± 1.9	50.0 ± 3.1	$\mathbf{61.9 \pm 1.0}$	$\mathbf{68.7 \pm 1.0}$	$\mathbf{66.3 \pm 0.7}$	38.1 ± 1.7	51.9 ± 3.1
$DA_{\lambda=0.50}$	61.2 ± 1.3	58.3 ± 2.0	50.2 ± 2.5	61.6 ± 0.5	67.4 ± 1.6	66.1 ± 0.9	37.3 ± 0.6	51.4 ± 5.2
$DA_{\lambda=0.75}$	59.7 ± 1.3	$\mathbf{58.6 \pm 3.4}$	51.8 ± 1.5	$\mathbf{61.9 \pm 1.3}$	66.8 ± 1.0	66.0 ± 0.1	38.2 ± 0.9	$\mathbf{54.7 \pm 2.3}$
$DA_{\lambda=1.00}$	60.8 ± 0.9	56.3 ± 1.8	48.3 ± 1.4	60.8 ± 0.3	67.1 ± 1.1	66.0 ± 0.7	37.6 ± 2.8	54.0 ± 2.2
$DA_{\lambda=10.00}$	58.7 ± 0.1	57.6 ± 1.7	52.1 ± 1.7	$\mathbf{61.9 \pm 1.5}$	65.9 ± 2.8	64.9 ± 0.8	38.4 ± 1.1	49.9 ± 5.4

Table 2: Weighted F1 scores for all subtasks, with results obtained through experiments with domain adaptation (DA) setup. Letters in brackets indicate the subtask. λ is an adversarial loss scaling factor. The DA approach outperforms the baseline in all cases, except for Amharic and Oromo, as indicated by the scores.

Language	mult.	or	tg
$DA_{\lambda=1.00}$	66.0	37.6	54.0
DA_{tok}	64.5	38.9	51.3
DA_{fam}	64.7	39.6	55.1

Table 3: Weighted F1 scores for additional experiments. The differences between models only in discriminator input for DA_{tok} and domain classes for DA_{fam}.

procedure for mapping languages and their families was done ad-hoc, which could affect the results. Further investigations with a more careful mapping and selection of languages for fine-tuning would be interesting for investigation.

6.4 Discriminator Input

Comparison of the systems with sequence and token level discriminator input show that a discriminator operating at the token level performs less favorably compared to one that operates at the sequence level. This observation leads us to formulate a hypothesis: a discriminator operating in a space of semantic meaning would help in achieving better performance.

6.5 Rankings Comparison

The average rank we achieved on all tracks is 20.4 with the highest rank of 9 for Amharic. The results were obtained with a single model trained on the Subtask B data with all the languages. A small version of Afro-XLM-Roberta was used as a model for fine-tuning. Other Afro-XLM-Roberta model sizes as well as other models adapted to African languages were not considered in our study.

7 Conclusion

In conclusion, our paper proposed a language-centric domain adaptation approach for sentiment analysis of low-resource African languages in SemEval-2023 Task 12. Our method utilized adversarial training with a small version of Afro-XLM-Roberta as the generator model, resulting in an improvement of up to 4.3 points in weighted F1 compared to the baseline while maintaining computational efficiency. Additionally, our findings suggested that using language family-based labels in adversarial training can enhance the quality of the cross-lingual setup for unseen languages. We also demonstrated that training the model on closely related languages in the language family tree can improve the quality of sentiment analysis for low-resource languages using language-centric domain adaptation. Our approach has the potential to improve sentiment analysis for African languages and other low-resource languages.

8 Limitations

In this section, we would like to emphasize several limitations of our study. First, the language mapping to language families was done ad-hoc. Other ways to map languages, as well as comparison of performance for different mappings, are kept for future experiments. Second, the discriminator used in our experiments was limited to a feed-forward network. As an alternative, other models such as RNN or transformer decoder could be explored in future studies. Third, our experiments were conducted using a single model, and we did not investi-

gate the behavior of other Afro-XLM-Roberta sizes or other models. Finally, language script was not taken into account in our study. Overall, the listed limitations suggest that further research is needed to fully understand the aspects of domain adaptation application to the domains that are based on language information.

9 Acknowledgement

We would like to express our gratitude to the organizers of SemEval-2023 Task 12 for providing us with the opportunity to participate in this exciting challenge. Their efforts have contributed to advancing research in the field of sentiment analysis for low-resource African languages.

References

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Kia Dashtipour, Soujanya Poria, Amir Hussain, Erik Cambria, Ahmad Y. A. Hawalah, Alexander Gelbukh, and Qiang Zhou. 2016. Multilingual sentiment analysis: State of the art and independent comparison of techniques. *Cognitive Computation*, 8(4):757–771.

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.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2020. Multi-source domain adaptation for text classification via distancenet-bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7830–7838.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Steven Arthur. 2023a. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif M. Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023b. SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

Subhabrata Mukherjee and Pushpak Bhattacharyya. 2013. Sentiment analysis : A literature survey.

Finn Årup Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs.

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 *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

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

Orith Toledo-Ronen, Matan Orbach, Yoav Katz, and Noam Slonim. 2022. Multi-domain targeted sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2751–2762, Seattle, United States. Association for Computational Linguistics.

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. 2018. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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*, volume 32. Curran Associates, Inc.