MSMO-ABSA: Multi-Scale and Multi-Objective Optimization for Cross-Lingual Aspect-Based Sentiment Analysis

Chengyan Wu*[1], Bolei Ma*[2], Ningyuan Deng[3], Yanqing He‡[3], Yun Xue[1], Xiaoyong Liu‡[4,5]

[1]South China Normal University, Foshan, China
[2]LMU Munich & Munich Center for Machine Learning, Munich, Germany
[3]Institute of Scientific and Technical Information of China, Beijing, China
[4]Guangdong Polytechnic Normal University, Heyuan, China
[5]Jilin Engineering Normal University, Changchun, China

{chengyan.wu,xueyun}@m.scnu.edu.cn, bolei.ma@lmu.de, {dengny2022,heyq}@istic.ac.cn, liuxy@gpnu.edu.cn

Abstract

Aspect-based sentiment analysis (ABSA) garnered growing research interest in multilingual contexts in the past. However, the majority of the studies lack more robust feature alignment and finer aspect-level alignment. In this paper, we propose a novel framework, **MSMO**: **M**ulti-**S**cale and **M**ulti-**O**bjective optimization for cross-lingual ABSA. During multi-scale alignment, we achieve cross-lingual sentence-level and aspect-level alignment, aligning features of aspect terms in different contextual environments. Specifically, we introduce code-switched bilingual sentences into the language discriminator and consistency training modules to enhance the model's robustness. During multi-objective optimization, we design two optimization objectives: supervised training and consistency training, aiming to enhance cross-lingual semantic alignment. To further improve model performance, we incorporate distilled knowledge of the target language into the model. Results show that MSMO significantly enhances cross-lingual ABSA by achieving state-of-the-art performance across multiple languages and models. [1]

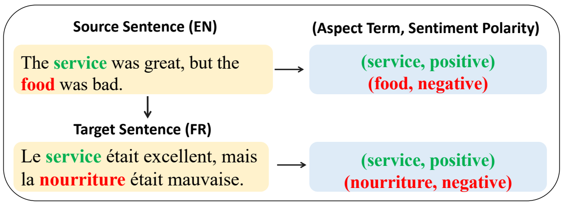

Figure 1: An example of a cross-lingual ABSA task. We train on the source language and perform aspect term extraction and sentiment polarity prediction on the target language.

1 Introduction

Aspect-based sentiment analysis (ABSA) involves identifying specific aspect terms and their sentiment polarity within a sentence (Liu, 2012; Pontiki et al., 2014). While research in ABSA has seen success with English texts, real-world social media interactions often involve multiple languages (Mao et al., 2022; Zhang et al., 2021b), highlighting the need for cross-lingual sentiment analysis. For example, as illustrated in Figure 1, if English is the source language and French is the target, a model trained on a labeled dataset in English should be able to identify the aspect terms "service" and "nourriture" in the French sentence, with sentiments "positive" and "negative" respectively.

Since obtaining large amounts of annotated training data for low-resource languages is extremely expensive, early cross-lingual sentiment analysis efforts (Zhou et al., 2016; Xu and Wan, 2017; Barnes et al., 2018) solely rely on annotated data from different source languages to learn sentiment classification for target languages. These models typically depend on bilingual dictionaries, pre-trained cross-lingual word embeddings, or machine translation to bridge the gap between source and target languages.

With the advent of multilingual pre-trained language models, recent research has shifted focus to data-level alignment, leveraging multilingual pre-trained models to fine-tune aligned translated data to bridge the gap between source and target languages (Li et al., 2021; Zhang et al., 2021a; Bigoulaeva et al., 2023; Lin et al., 2023).

To further enhance the awareness of the multilingual ABSA task, and inspired by previous adversarial training (Wang and Pan, 2018; Zhou et al., 2022) and consistency training methods (Wang and Henao, 2021; Zhou et al., 2022), we propose a framework based on **m**ulti-**s**cale and **m**ulti-**o**bjective optimization, called **MSMO**. Specifically, the MSMO framework comprises four key

*Equal contributions.
‡Corresponding authors.
[1]Code at https://github.com/swaggy66/MSMO

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 18723–18735

components: a feature extractor, a language discriminator, a consistency training module, and a sentiment classifier. For the multi-scale aspect, we employ adversarial training for sentence-level alignment and consistency training for aspect-level alignment, leveraging both the bilingual translated dataset and the code-switched dataset in the process. Specifically, introducing a code-switched dataset that switches different aspect terms can introduce perturbations, allowing the embedding spaces of the source and target languages to align better with the anchor aspects and improving the robustness of the model. For the multi-objective optimization, we combine supervised training and consistency training as optimization objectives, aiming to align aspect terms of different languages at a finer granularity. Additionally, we extend the MSMO framework to multilingual ABSA. To explore the importance of unlabeled target language knowledge for performance improvement, we also apply knowledge distillation using unlabeled data in the target language.

In summary, our main contributions are:

- **Sentence-level Alignment**: We propose an adversarial training approach using a code-switched dataset. This method enhances the language discriminator's ability to capture invariant features and develop more robust representations across languages by introducing aspect term perturbations.
- **Aspect-level Alignment**: We utilize consistency training to ensure the model provides consistency predictions for aspect terms with the same sentiment, improving alignment at the aspect level.
- **Multi-objective Optimization**: We integrate supervised training and consistency training objectives to minimize the performance gap between different languages.
- **Extensive Evaluation**: We conduct comprehensive experiments on benchmark datasets in five languages across cross-lingual and multilingual settings. The results demonstrate that our MSMO achieves state-of-the-art performance.

2 MSMO Framework

2.1 Problem Formulation and Background

We regard the ABSA task as a sequence labeling problem. Given an input sentence $x = \{x_1, x_2, \ldots, x_n\}$ containing n tokens, our goal is to predict the labels $y = \{y_i\}_{i=1}^{n}$ for the input sequence, where $y_i \in Y = \{B, I, E, S\}$-$\{POS, NEU, NEG\} \cup \{O\}$, representing the aspect term boundaries and their sentiment polarities corresponding to the token x_i.

In our cross-lingual transfer framework, following Zhang et al. (2021a), we use the source language data D_S and the translated target language data D_T. We also use the code-switched data (D_{S_T}, D_{T_S}) during training, where D_{S_T} is created by replacing the aspect terms in D_S with their counterparts in the target language, and D_{T_S} is created by replacing the aspect terms in D_T with their counterparts in the source language. The training data D_U consists of sentence-label pairs $(x_u, y_u) \in D_U$, where $D_S \cup D_T \cup D_{S_T} \cup D_{T_S} \in D_U$, aiming to predict the label sequence y_t for the target language in the test set.

2.2 Preliminaries of the MSMO Framework

Figure 2 illustrates the key components of our method, which mainly consists of two overall steps: (1). Sentence-level alignment by adversarial training (§2.3); (2). Aspect-level alignment with multi-objective optimization (§2.4). Before delving into the details of the two steps, we illustrate the preliminaries of the Pretrained Multi-Lingual Encoder.

The MSMO framework integrates a pre-trained multilingual encoder (M) as a feature extractor to generate contextual representations of sentence tokens. Given a sequence of n tokens $[x_1, x_2, \ldots, x_n]$, we take the final hidden layer outputs of the M as the intermediate representations $h_i \in \mathbb{R}^l$.

$$[h_1, h_2, \ldots, h_n] = M([x_1, x_2, \ldots, x_n]) \quad (1)$$

These intermediate token embeddings h_i are then fed into two different branches. In the first stage (§2.3), h_i is input into the language discriminator Q, which aims to predict a scalar score indicating whether x is from the source or the target. In the second stage (§2.4), we use the updated M from the first stage to encode and input the representations into the sentiment classifier P and the consistency training module C. Q uses a sigmoid activation function, while both P and C use softmax activation functions, defined as:

$$p(x_i) = F(\text{Dropout}(W * h_i + b)) \quad (2)$$

P predicts labels for the input sequences based on the feature representation h_i, while C aligns aspect terms with the same sentiment polarity across

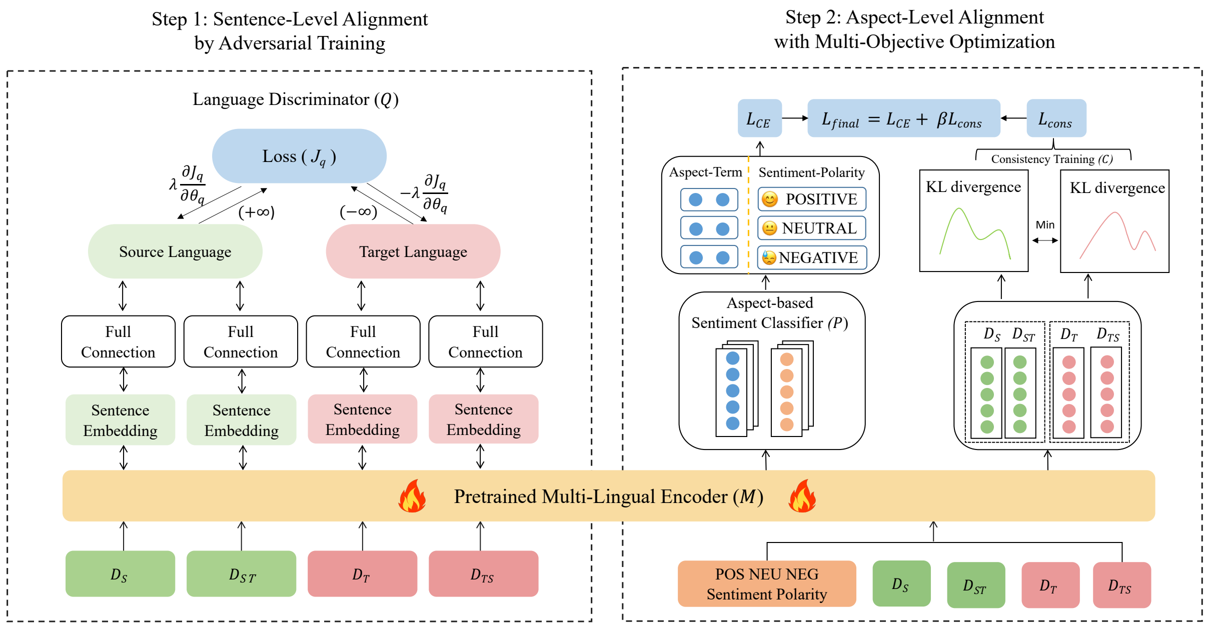

Figure 2: The MSMO framework. It mainly comprises two basic steps: (1). Sentence-level alignment by adversarial training (§2.3); (2). Aspect-level alignment with multi-objective optimization (§2.4). The Pretrained Multi-Lingual Encoder connects both steps by updating the parameters from the loss of language discriminator in step 1 and from the combined loss in step 2.

different languages by aligning the predicted probability distributions of aspect terms in the source language and target language.

2.3 Step 1: Sentence-Level Alignment by Adversarial Training

Now we look into the first step which leverages a language discriminator to conduct sentence-level alignment by adversarial training. In previous research, ADAN (Chen et al., 2018) is proposed to use the Wasserstein distance (Arjovsky et al., 2017) for standard adversarial training, addressing the instability of adversarial training in the ADAN-GRL (Ganin et al., 2016) method and achieving better performance. Inspired by the ADAN paradigm, we design a language discriminator Q to reduce the semantic gap across languages. Q is a binary classifier with a sigmoid layer on top, so the language recognition score is always between 0 and 1, aiming to determine the probability that the input text x is from source or target based on the hidden features h_i captured by the feature extractor. For training, Q is connected to the encoder via a gradient reversal layer (Ganin and Lempitsky, 2015), which retains the input during the forward pass but multiplies the gradient by $-\lambda$ during the backward pass (we set λ to 1 in Figure 2). Thus, standard backpropagation can be used to train the entire network holistically. However, unlike ADAN, to enhance the robustness of the model while preserving semantic consistency, we also introduce the code-switched dataset in addition to using bilingual parallel corpora. Specifically, for source language sentences that introduce aspect terms in the target language, we stimulate the language discriminator Q to identify them as source language sentences, and *vice versa*.

By inducing local perturbations from aspect term changes in sentences, we encourage the language discriminator to recognize common features of languages and stimulate the feature extractor to capture invariant features of languages through backpropagation. Our objective is to approximately minimize the Wasserstein distance between $(P(h_i), P(h_i'))$ according to the Kantorovich-Rubinstein duality (Villani, 2013). To ensure that Q is a Lipschitz function (up to a constant), the parameters of Q are always clipped to a fixed range $[-c, c]$. Thus, the objective J_q of Q becomes:

$$J_q(P(h_i), P(h_i')) \equiv \max_{\theta_q} \mathbb{E}[\mathcal{Q}(\mathcal{P}(h_i))] - \mathbb{E}[\mathcal{Q}(\mathcal{P}(h_i'))] \tag{3}$$

where $h_i \in D_S \cup D_{S_T}$, and $h_i' \in D_T \cup D_{T_S}$. The supremum (maximum) of J_q is taken over the set of all 1-Lipschitz functions Q. Intuitively, Q tries to output higher scores for source instances and lower scores for target instances. More formally, J_q is an approximation of the Wasserstein distance between $P(h_i)$ and $P(h_i')$ in Equation 3.

2.4 Step 2: Aspect-Level Alignment with Multi-Objective Optimization

After the first stage of training, the updated encoder has learned to extract invariant features across different languages. We use the updated encoder for the second stage of training, where the extracted features are fed into the sentiment classifier P and the consistency training module C.

Supervised Training. For the sentiment classifier P, we use the traditional cross-entropy loss, denoted as $\mathcal{L}_{CE}$, which is computed between the predicted label distribution and the gold label in one-hot encoding. Therefore, we seek to minimize the following loss function for P:

$$\mathcal{L}_{CE} = \frac{1}{|D_U|} \sum_{(\boldsymbol{x},\boldsymbol{y}) \in D_U} \left[-\frac{1}{L} \sum_{i=1}^{L} y_i \log p_\theta \left(y_i \mid x_i \right) \right] \quad (4)$$

where L represents the length of the sentence X, i denotes the i-th token in the sentence. In addition, (x, y) belongs to the labeled training dataset D_U.

Consistency Training. Consistency training (Miyato et al., 2019; Clark et al., 2018; Xie et al., 2020) aims to reduce the model's overfitting and bias towards specific input forms by guiding the model to produce consistency predictions under different input perturbations. Although it has been successful in CV (Wang et al., 2024a) and sentence-level NLP tasks (Miyato et al., 2019; Xie et al., 2020), there has been a lack of effective attempts at consistency training for the cross-lingual ABSA task. In cross-lingual ABSA, learning aspect-level feature representations across different languages can achieve cross-lingual adaptability from a finer-grained perspective, where the model should maintain consistency in the predictions of aspect terms with the same sentiment polarity.

We explore a consistency training method for cross-lingual ABSA to improve alignment at the aspect level, as follows. Let ϕ be a transformation function that generates small perturbations, such as noise from text translation or aspect term transformation. In this paper, one transformation method is to translate X into another language, and another is to swap the aspect terms in the source and target language sequences. Given a sequence of tokens X and aspect terms in the sequence denoted as s, we apply the ϕ transformation to the source language sequence to obtain a perturbed sequence X', where the aspect terms s after transformation correspond to s'. We encourage the model to capture feature representations of aspect terms with the same sentiment across different languages and use bidirectional KL divergence to compute the divergence D_{div} between the probability distributions of the aspect term pairs (s, s') at the span level, and then minimize their consistency loss $\mathcal{L}_{\text{cons}}$ to output consistent probability distributions over s and s':

$$\mathcal{L}_{\text{cons}} = \frac{1}{m} \sum_{(s_i, s_i') \in (\mathbf{X}, \mathbf{X}')} \frac{1}{2} \left[\mathrm{KL}\left(P\left(y_i' \mid s_i'\right) \| P\left(y_i \mid s_i\right) \right) + \mathrm{KL}\left(P\left(y_i \mid s_i\right) \| P\left(y_i' \mid s_i'\right) \right) \right] \quad (5)$$

where y_i and y_i' are the labels of the spans and m is the total number of aspect term pairs (s_i, s_i'). Inspired by the approach of Zhou et al. (2022), we define the probability of a span as the product of the tokens that constitute the span.

Multi-Objective Optimization. As described above, we apply supervised training and consistency training to the source language data, translated target language data, and code-switched data. Then, we combine the cross-entropy loss $\mathcal{L}_{\text{CE}}$ with the consistency loss $\mathcal{L}_{\text{cons}}$ to form our total training objective:

$$\mathcal{L}_{\text{total}} = \sum_{X \in D_{\text{U}}} \mathcal{L}_{\text{CE}} + \sum_{X \in D_{\text{U}}} \beta \mathcal{L}_{\text{cons}} \quad (6)$$

3 Experimental Setups

3.1 Dataset

We choose the SemEval-2016 dataset (Pontiki et al., 2016) to evaluate our method. This dataset consists of real user reviews across eight languages, with ABSA annotations available for English (EN), French (FR), Spanish (ES), Dutch (NL), Russian (RU), and Turkish (TK). However, due to the limited size of the Turkish test set (fewer than 150 sentences), we excluded it from our evaluation, consistent with prior multilingual ABSA studies (Zhang et al., 2021a; Lin et al., 2023). For a fair comparison, we use the data processed by Zhang et al. (2021a). The data for each language is divided into training, validation, and test sets, along with a code-switched dataset. We use English as the source language during training, and the other languages as target languages in the prediction phase.

We utilize the code-switched dataset proposed by Zhang et al. (2021a), referred to as D_{S_T} and D_{T_S}. D_{S_T} is constructed by replacing the aspect terms in the source language dataset D_S with aspect terms that appear in the target language dataset D_T. In contrast, D_{T_S} is generated by replacing the aspect terms in the target language dataset D_T with aspect terms that appear in the source dataset D_S.

Figure 3 shows examples of the source language, translated, and code-switched datasets. The aspect terms “service” and “food” in the source language sentence D_S are marked with special symbols (e.g., “{ }”, “[]”). The corresponding target language sentence D_T is obtained through machine translation, with the aspect terms being “service” and “la nourriture”. Then, by matching the special symbol markers, we can swap the aspect terms between the source and target language sentences. That is, “service” and “food” in the source language sentence is replaced with “service” and “la nourriture” in the target language to obtain D_{S_T}, and vice versa for the target language sentence to obtain D_{T_s}, thereby generating the code-switched data.

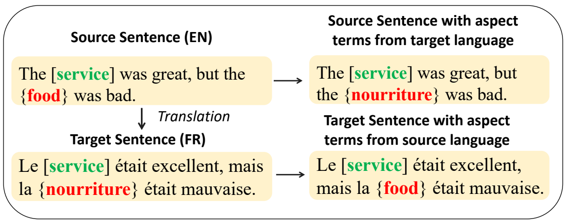

Figure 3: An example of the code-switched dataset.

The details of the dataset statistics in each language are shown in Table 1. # S and # A denote the number of sentences and aspect terms in different sets, respectively.

		EN	FR	ES	NL	RU
Train	# S	2000	1664	2070	1722	3655
	# A	1743	1641	1856	1231	3077
Test	# S	676	668	881	575	1209
	# A	612	650	713	373	949

Table 1: Statistic of the original dataset.

3.2 Models and Parameter Settings

We evaluate four target languages using the Micro-F1 metric on two multilingual pre-trained models, including the cased mBERT (Devlin et al., 2019) and the base XLM-R model (Conneau et al., 2020), for a fair comparison with existing methods. A prediction is considered correct only when the tuple (entity, label) is correctly predicted, where the entity is the boundary of the aspect term and the label is the corresponding sentiment polarity. Following the settings of Zhang et al. (2021a), we set the maximum training steps to 2000 for mBERT and 2500 for XLM-R. Additionally, we allocate different weights for the multi-objective optimization functions of the two models across the four target languages, as in Equation 6, with β values {4.5e-4, 2.5e-4, 2.5e-4, 3.5e-4} and {2.5e-3, 1.5e-3, 1.5e-3, 3.5e-3} respectively (refer to Appendix 4.4). Based on the performance of the validation set of the source language, we select the best model in the last 500 steps.

We select the optimal training hyperparameters through a grid search over combinations of batch size and learning rate. The ranges are: learning rate {1e-5, 2e-5, 5e-5}; batch size {8, 16, 25}. For mBERT, we use a learning rate of 5e-5 and a batch size of 16; for XLM-R, we use a learning rate of 2e-5 and a batch size of 8. For all experiments, we report the average F1 scores over 5 runs with different random seeds.

The MSMO method integrates several advanced components, such as multi-teacher distillation and consistency training, which introduce additional computational demands during both training and inference. To quantify this overhead, we report the resources used in our experiments. All experiments were run on a single NVIDIA A6000 GPU (48 GB memory), a CPU with 120 MB ($\approx$ 0.12 GB) of L3 cache, and 42 GB of RAM.

Model	**Variant**	**GPU Memory**
mBERT	MSMO	≈22
	w/o Language Discriminator	≈18 GB
	w/o Consistency Training	≈16 GB
XLM-R	MSMO	≈27
	w/o Language Discriminator	≈24 GB
	w/o Consistency Training	≈21 GB

Table 2: Approximate GPU memory usage.

3.3 Knowledge Distillation Settings

Figure 4 and 5 illustrate the three modes of knowledge distillation: single-teacher distillation, multi-teacher distillation, and multilingual distillation, proposed by Zhang et al. (2021a).

For the teacher model, we use a combination of source and target language data (D_T, D_S, D_{S_T}, and D_{T_S}) to train the teacher model within our

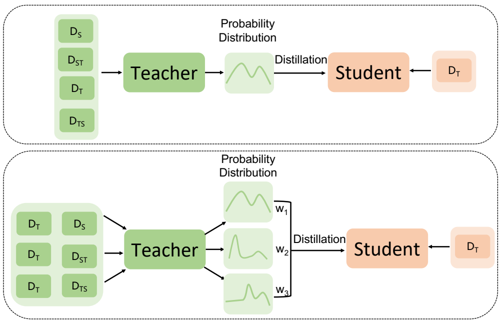

Figure 4: The single-teacher and multi-teacher distillation process.

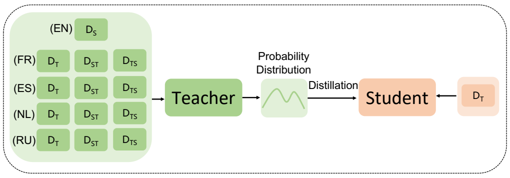

Figure 5: The multilingual distillation process.

MSMO framework. For the student model, to enable it to perform the ABSA task in the target language, we use the translated target language data as the initial training data, then obtain soft labels for the target language test set predictions from the teacher model, and finally conduct incremental training on this soft-labeled data. For multi-teacher distillation, we assign equal weights to different teacher models, i.e., $w_k = 1/3$ in Equation 7.

$$p_t = \sum_{k=1}^{3} \omega_k * g_{t_k} \tag{7}$$

where w_k is the weight for each teacher model. With the combined soft label g_t, a student model can be trained similarly by using only the encoder and sentiment classifier modules as in Equation 8.

$$\mathcal{L}_{KD} = \frac{1}{|D_{NL}|} \sum_{X \in D_{NL}} \left[\frac{1}{L} \sum_{i=1}^{L} \text{MSE}\left(p_{t_i}, p_{s_i}\right) \right] \tag{8}$$

where L represents the length of the sentence X, and i denotes the i-th token in the sentence. Additionally, D_{NL} indicates the unlabeled dataset in the target language, and p_{t_i} and p_{s_i} are the prediction probabilities of the i-th token from the student and teacher models, respectively. We use the mean squared error loss $\text{MSE}(\cdot)$ to measure the difference between the two probability distributions.

3.4 Compared Methods

We compare our method against several baseline approaches:

- **SUPERVISED**: A fully supervised method where the model is trained using data from the target language.
- **ZERO-SHOT** (Conneau et al., 2020): This method fine-tunes the model on labeled source data and applies it directly to target data, showing strong cross-lingual adaptation.
- **TRANSLATION-TA** and **BILINGUAL-TA** (Li et al., 2021): Translation-based. TRANSLATION-TA trains the model with pseudo-labeled data and aligned translations, while BILINGUAL-TA combines source data with aligned translations for training.
- **ACS** (Zhang et al., 2021a): This method introduces code-switched data for training, aligning aspect terms at the data level to bridge the gap between different languages.
- **CL-XABSA** (Lin et al., 2023): It performs contrastive learning at both the sentiment levels (SL) and token levels (TL), respectively, minimizing the distance between tokens with the same sentiment and identical labels.
- **Equi-XABSA** (Lin et al., 2024): It mitigates sample imbalance and languages' representation disparity.

Additionally, we evaluate the performance of our method with distilled data compared to:

- **Single-Teacher Distillation**: ACS-DISTILL-S and CL-XABSA-DISTILL-S.
- **Multi-Teacher Distillation**: ACS-DISTILL-M and CL-XABSA-DISTILL-M.

4 Results and Analysis

4.1 Cross-lingual ABSA Results

We compare our method with previous methods in Table 3. Overall, we achieve SOTA performance on both mBERT and XLM-R models compared to the zero-shot baselines and the CL-XABSA baselines from Lin et al. (2023), indicating that our approach better facilitates semantic convergence across different languages. Additionally, our results are closer to those of the fully supervised fine-tuning method, highlighting the robustness and effectiveness of our approach in bridging the performance gap with supervised methods.

Content-wise, we observe the following in-depth key phenomena:

XLM-R vs. mBERT. Methods based on the XLM-R backbone generally outperform those based on the mBERT backbone. The primary rea-

Methods	mBERT FR	ES	NL	RU	Avg	XLM-R FR	ES	NL	RU	Avg
SUPERVISED	61.80	67.88	56.80	58.87	61.34	67.44	71.93	64.28	64.93	67.15
ZERO-SHOT (Conneau et al., 2020)	45.60	57.32	42.68	36.01	45.40	56.43	67.10	59.03	56.80	59.84
TRANSLATION-TA (Li et al., 2021)	40.76	50.74	47.13	41.67	45.08	47.00	58.10	56.19	50.34	52.91
BILINGUAL-TA (Li et al., 2021)	41.00	51.23	49.72	43.67	46.41	49.34	61.87	58.64	52.89	55.69
ACS (Zhang et al., 2021a)	49.65	59.99	51.19	52.09	53.23	59.39	67.32	62.83	60.81	62.59
CL-XABSA (SL) (Lin et al., 2023)	49.75	60.12	49.34	50.10	52.32	58.10	64.85	59.75	58.84	60.39
CL-XABSA (TL) (Lin et al., 2023)	50.55	60.09	52.45	50.73	53.46	59.47	64.63	59.40	61.13	61.16
Equi-XABSA (Lin et al., 2024)	50.08	63.08	51.85	52.59	54.40	60.68	69.56	61.31	62.34	63.47
MSMO	**51.42**	**63.26**	**52.68**	**53.45**	**55.20**	**61.01**	**69.74**	**63.26**	**62.52**	**64.13**
ACS-DISTILL-S (Zhang et al., 2021a)	52.23	62.04	52.72	53.00	55.00	61.00	68.93	62.89	60.97	63.45
ACS-DISTILL-M (Zhang et al., 2021a)	52.25	62.91	53.40	54.58	55.79	59.90	69.24	63.74	62.02	63.73
CL-XABSA-DISTILL-S (Lin et al., 2023)	52.76	62.54	53.38	53.48	55.27	61.20	69.13	63.01	61.37	63.68
CL-XABSA-DISTILL-M (Lin et al., 2023)	52.99	63.54	53.52	53.98	56.01	62.10	69.37	64.27	62.29	64.51
MSMO-DISTILL-S	**53.58**	**63.80**	**53.97**	**54.47**	**56.46**	**61.69**	**70.16**	**63.58**	**62.96**	**64.25**
MSMO-DISTILL-M	**54.39**	**64.59**	**54.14**	**54.89**	**56.94**	**63.89**	**69.93**	**65.15**	**63.20**	**65.54**

Table 3: Performance comparison of various methods on different languages using mBERT and XLM-R. S denotes single-teacher distillation and M denotes multi-teacher distillation.

son is that XLM-R has a larger number of parameters and uses a larger multilingual corpus during the pre-training phase, leading to stronger cross-lingual adaptation capabilities.

Performance Comparison. Our proposed MSMO method not only outperforms the ZERO-SHOT method and translation-based methods (BILINGUAL-TA and TRANSLATION-AF) but also achieves better performance than the CL-XABSA method. This demonstrates the effectiveness of introducing sentence-level adversarial training and aspect-level alignment between different languages.

Language-Specific Improvements. Our MSMO method achieves performance improvements across all four target languages, with a more noticeable improvement in Spanish. Compared to the CL-XABSA method, the performance of MSMO in Spanish improves by 3.14% and 4.89% on mBERT and XLM-R, respectively. The primary reason is that languages from different families have different semantic spaces, and Spanish is closer to English in terms of language family, making their semantic spaces more easily converged. This aligns with the findings of Zhang et al. (2021a).

Distillation Performance. Following the ACS paradigm, we apply the MSMO method to single-teacher distillation (MSMO-DISTILL-S) and multi-teacher distillation (MSMO-DISTILL-M), and both achieve higher performance. This can be explained by the fact that the teacher model of our MSMO method outperforms the teacher model of the CL-XABSA method. During the knowledge distillation process to the target language, the teacher model of the MSMO method can achieve more accurate soft label predictions, thereby better guiding the student model to learn from these soft labels on unlabeled target language data. Additionally, the multi-teacher distillation method performs better than the single-teacher distillation method, which may be because the multi-teacher model can better combine the strengths of different teachers to guide the student model.

4.2 Multilingual ABSA Results

To fairly compare with the previous SOTA method by Lin et al. (2023), we report the results of the MSMO method in a multilingual setting (MTL-MSMO) in Table 4. As shown in Table 4, the MTL-MSMO method outperforms the multilingual CL-XABSA method (MTL-CL-XABSA) for the teacher model, with average Micro-F1 improvements of 1.61% and 0.77% on the mBERT and XLM-R models, respectively. This indicates that using the MSMO method for distillation with unlabelled data can achieve higher performance. This improvement can be attributed to our language discriminator and consistency training modules, which better align the semantic spaces in a multilingual setting. Furthermore, due to the superior performance of the teacher model, the student model can learn knowledge from multiple target languages through the soft labels predicted by the MTL-MSMO teacher model and apply it to specific language inference. Our method also shows significant improvement in multilingual knowledge distillation, with average Micro-F1 improvements of 0.97% and 0.77% on the mBERT and XLM-R models, respectively, over the previous MTL-CL-XABSA-DISTLL, demonstrating the effectiveness and superiority of our proposed method.

	FR	ES	NL	RU	Avg
Based on mBERT:					
MTL-CL-XABSA	50.01	59.05	51.22	50.59	52.72
MTL-CL-XABSA-DISTLL	53.03	62.19	54.25	54.63	56.03
MTL-MSMO	**51.33**	**60.43**	**53.68**	**51.89**	**54.33**
MTL-MSMO-DISTLL	**54.56**	**62.69**	**55.56**	**56.19**	**57.00**
Based on XLM-R:					
MTL-CL-XABSA	60.09	68.88	64.16	63.07	64.05
MTL-CL-XABSA-DISTILL	62.37	70.58	65.98	62.79	65.43
MTL-MSMO	**60.93**	**69.34**	**64.85**	**64.17**	**64.82**
MTL-MSMO-DISTLL	**63.23**	**70.95**	**66.24**	**64.36**	**66.20**

Table 4: Multilingual results (MTL) with mBERT and XLM-R as backbone respectively. MTL-CL-XABSA and MTL-CL-XABSA-DISTLL are selected from the best result between TL and SL in Lin et al. (2023).

4.3 Ablation Study

To demonstrate the effectiveness of the main components in the MSMO method, we experiment with ablations and present the results in Table 5. We design two variants of MSMO for the experiments:

1) **w/o. Language Discriminator**: We remove the language discriminator, retaining only the feature extractor, sentiment classifier, and consistency training modules to train the model;

2) **w/o. Consistency Training**: We remove the consistency training module, retaining only the feature extractor, sentiment classifier, and language discriminator to train the model.

As shown in Table 5, the results indicate that

1) MSMO experiences a performance drop when any component is removed, demonstrating the importance of all components;

2) Removing the language discriminator results in a decrease of 1.68% and 1.30% in the average F1 score, respectively, indicating that training the language discriminator with bilingual data and code-switched bilingual data helps the model learn language-invariant features, contributing to performance improvement. Specifically, in different code-switched contexts, the model can better focus on the boundary changes of aspect terms in different languages, thereby better learning aspect term boundaries and improving model robustness;

3) Removing the consistency training component results in a decrease of 1.57% and 1.08% in the average F1 score, respectively, indicating that narrowing the predicted probability distributions of aspect terms in different contexts can reduce the discrepancy of aspect terms with the same sentiment polarity in different semantic spaces.

	FR	ES	NL	RU	Avg
Based on mBERT:					
w/o. Language Discriminator	49.70	60.61	51.57	52.21	53.52
w/o. Consistency Training	50.59	60.40	51.30	52.25	53.63
MSMO	**51.42**	**63.26**	**52.68**	**53.45**	**55.20**
Based on XLM-R:					
w/o. Language Discriminator	59.82	68.10	62.41	60.99	62.83
w/o. Consistency Training	59.51	67.96	62.91	61.82	63.05
MSMO	**61.01**	**69.47**	**63.26**	**62.52**	**64.13**

Table 5: The ablation study results.

4.4 The Impact of the Parameter β

In Figure 6, we present the impact of the parameter β in Eq. 6 on the model performance. We observe the following phenomena: 1) Spanish achieves the best performance at lower values of β, which can be explained by the fact that similar languages share a closer semantic space and feature representations, making alignment easier. 2) Both excessively large and small values of β lead to performance degradation. When β is too small, the model may overly rely on supervised training, resulting in poor cross-lingual generalization. On the other hand, when β is too large, the model may focus too much on the consistency loss, neglecting language-specific details in sentiment analysis, which leads to suboptimal classification performance for certain languages. Therefore, the choice of β should be adjusted according to the specific requirements of the task and the characteristics of the data, ensuring that the model can capture cross-lingual consistency without losing sensitivity to language-specific features.

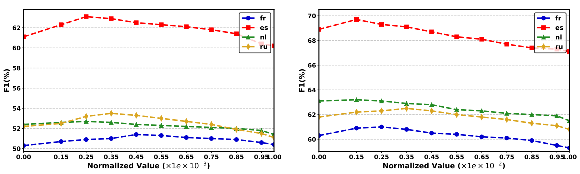

Figure 6: The impact of the parameter β. Left: mBERT performance. Right: XLM-R performance.

4.5 Excursion: Comparison with LLMs

As recent advances in LLMs emerge, we also evaluate the LLM performance on the ABSA task to compare with our method. We evaluate the LLMs in two ways: by zero-shot prompting (Wu et al., 2025b), and by fine-tuning the LLMs on the source language (English) and then validating on target languages (similar to the supervised baseline). For zero-shot prompting, we apply the instruction-tuned version of GPT-4o (OpenAI et al., 2024), Gemma-2 9B (Team et al., 2024), Llama-3.1 8B (Grattafiori et al., 2024), and Mistral 7B (Jiang et al., 2023), and Qwen2.5 7B (Qwen et al., 2025). For fine-tuning, we employ the LoRA approach (Hu et al., 2022) on the four open-weight LLMs using the base models. We present the results in Table 6. We apply the same prompt used in Wu et al. (2025b) to the models in zero-shot settings with only the prompt template as the system prompt.

Still, our proposed MSMO method with multilingual distillation demonstrates a marked advancement in average performance, significantly outperforming the highest-scoring LLMs in zero-shot (GPT-4o) and in LoRA fine-tuning (Qwen-2.5). This suggests that while these LLMs excel in various NLP tasks, they may not be as effective for nuanced token-level classification tasks without additional fine-tuning (Wang et al., 2025; Nie et al., 2025). Notably, performance across languages varies, with Spanish achieving relatively higher scores and French lower compared to other languages. This observation is consistent with the results obtained from our MSMO approach.

	FR	ES	NL	RU	Avg
GPT-4o	48.43	49.91	49.94	45.15	48.36
Gemma-2-9b-It	50.94	48.80	50.24	39.34	47.33
Llama-3.1-8B-Instruct	23.15	32.49	33.53	30.18	29.84
Mistral-7B-Instruct-v0.3	37.21	38.32	33.98	26.58	34.02
Qwen2.5-7B-Instruct	48.88	48.29	46.75	40.25	46.04
Gemma-2-9B + LoRA	48.17	57.46	51.97	47.65	51.31
Llama-3.1-8B + LoRA	52.65	55.37	50.37	48.12	51.63
Mistral-7B-v0.3 + LoRA	59.46	60.79	56.89	50.52	56.92
Qwen2.5-7B + LoRA	63.01	68.95	60.84	53.50	61.58
MTL-MSMO-DISTLL on mBERT	**54.56**	**62.69**	**55.56**	**56.19**	**57.00**
MTL-MSMO-DISTLL on XLM-R	**63.23**	**70.95**	**66.24**	**64.36**	**66.20**

Table 6: Performance of different LLMs in zero-shot-prompting and in LoRA fine-tuning in comparison with MTL-MSMO-DISTLL (ours).

5 Related Work

Cross-Lingual ABSA. Research in cross-lingual ABSA (XABSA) generally falls into two categories: data alignment and embedding learning. Data alignment aims to incorporate language-specific knowledge into the target language, often using translation systems or dictionaries to convert annotated data from high-resource languages (Zhou et al., 2013). Techniques such as co-training (Zhou et al., 2015) and constrained SMT (Lambert, 2015) improve data quality. Additionally, pre-trained multilingual word embeddings and methods like warm-up mechanisms (Li et al., 2021) and shared vector spaces (Jebbara and Cimiano, 2019) enhance model performance in multi-lingual ABSA, with additional CNN-based architectures offering further gains (Wang et al., 2024b). Zhang et al. (2021a) adopt the translation-based methods by code-switching the aspect terms in target and source data for cross-lingual ABSA. Following this, Lin et al. (2023) use contrastive learning for cross-lingual ABSA. There is also newly publicized multilingual ABSA datasets for implicit aspects (Wu et al., 2025a).

Adversarial Networks. Adversarial training, popular in computer vision (Knoester et al., 2022), has seen limited application in ABSA. Notable work includes Miyato et al. (2017), who apply domain adversarial training to ABSA, and Wang and Pan (2018), who use adversarial networks to align feature vectors across languages. Some methods have also explored character and word-level perturbations. Mamta and Ekbal (2022) generate adversarial samples for specific aspects while maintaining semantic coherence. Adversarial training helps evaluate model resilience and identify vulnerabilities (Lin et al., 2023).

Consistency Training. Consistency training regularizes a model by ensuring predictions remain similar for both original and perturbed inputs (Zhou et al., 2022). While widely used in NLP, its application in ABSA is still emerging. Existing work includes Chen et al. (2022), which demonstrates that simple augmentations combined with consistency training yield competitive ABSA performance. Additionally, Zhang et al. (2023) introduces a sentiment consistency regularizer to maintain sentiment consistency across spans.

6 Conclusion

In this work, we introduce the novel application of adversarial training and consistency training to cross-lingual aspect-based sentiment analysis. Our approach includes language discriminator and consistency training modules at the sentence and aspect levels, respectively, to better align aspect terms across languages. Multi-objective optimization further bridges semantic gaps between languages, establishing a robust baseline. Additionally, we demonstrate the effectiveness of knowledge distillation with the MSMO method. Extensive experiments confirm that our approach outperforms previous state-of-the-art methods. Future work will explore extending the MSMO framework to other multilingual NLP tasks.

Limitations

Compared to the traditional cross-lingual ABSA methods, our proposed MSMO method incorporates different modules designed to learn the boundary features of aspect terms across different languages. However, the consistency of these features in highly diverse or idiomatic expressions may still present challenges, necessitating further refinement of these modules to handle more nuanced language variations. Additionally, our experiments rely on specific benchmark datasets, and whether our method can be generalized to other multilingual NLP tasks or real-world applications remains to be verified. Future work should include broader multilingual datasets to assess the robustness of our approach. We leave these for our future research.

Ethical Considerations

This research was conducted in accordance with the ACM Code of Ethics. The dataset (Pontiki et al., 2016) that we use is publicly available. We report only aggregated results in the main paper. We have not intended or do not intend to share any Personally Identifiable Data with this paper.

Acknowledgments

The authors acknowledge the use of ChatGPT exclusively to refine the text in the final manuscript and to assist in coding. CW and YX are supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515011370, the National Natural Science Foundation of China (32371114), the Characteristic Innovation Projects of Guangdong Colleges and Universities (No. 2018KTSCX049), and the Guangdong Provincial Key Laboratory (No. 2023B1212060076). ND and YH are supported in part by the National Science and Technology Major Project (2023ZD0121502). XL was supported by Project for Improving Scientific Research Capacity of Key Construction Disciplines in Guangdong(No. 2025ZDJS022), Project of Teaching Quality and Teaching Reform in Undergraduate Universities of Guangdong Province: Modern Industrial College of Smart Agriculture in GPNU, Project of Guangdong Polytechnic Normal University (Grant No.22GPNUZDJS16).

References

Martín Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR.

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.

Irina Bigoulaeva, Viktor Hangya, Iryna Gurevych, and Alexander Fraser. 2023. Label modification and bootstrapping for zero-shot cross-lingual hate speech detection. *Language Resources and Evaluation*, 57(4):1515–1546.

David Z. Chen, Adam Faulkner, and Sahil Badyal. 2022. Unsupervised data augmentation for aspect based sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6746–6751, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1914–1925, Brussels, Belgium. 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. 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.

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.

Yaroslav Ganin and Victor S. Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1180–1189. JMLR.org.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Soufian Jebbara and Philipp Cimiano. 2019. Zero-shot cross-lingual opinion target extraction. 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 2486–2495, Minneapolis, Minnesota. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Joris Knoester, Flavius Frasincar, and Maria Mihaela Truşcundefined. 2022. Domain adversarial training for aspect-based sentiment analysis. In *Web Information Systems Engineering – WISE 2022: 23rd International Conference, Biarritz, France, November 1–3, 2022, Proceedings*, page 21–37, Berlin, Heidelberg. Springer-Verlag.

Patrik Lambert. 2015. Aspect-level cross-lingual sentiment classification with constrained SMT. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 781–787, Beijing, China. Association for Computational Linguistics.

Xin Li, Lidong Bing, Wenxuan Zhang, Zheng Li, and Wai Lam. 2021. Unsupervised cross-lingual adaptation for sequence tagging and beyond. *Preprint*, arXiv:2010.12405.

Nankai Lin, Yingwen Fu, Xiaotian Lin, Dong Zhou, Aimin Yang, and Shengyi Jiang. 2023. Cl-xabsa: Contrastive learning for cross-lingual aspect-based sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2935–2946.

Nankai Lin, Meiyu Zeng, Xingming Liao, Weizhong Liu, Aimin Yang, and Dong Zhou. 2024. Addressing class-imbalance challenges in cross-lingual aspect-based sentiment analysis: Dynamic weighted loss and anti-decoupling. *Expert Systems with Applications*, 257:125059.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Mamta and Asif Ekbal. 2022. Adversarial sample generation for aspect based sentiment classification. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 478–492, Online only. Association for Computational Linguistics.

Yue Mao, Yi Shen, Chao Yu, and Longjun Cai. 2022. A joint training dual-mrc framework for aspect based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, page 13543–13551.

Takeru Miyato, Shin ichi Maeda, Masanori Koyama, and Shin Ishii. 2017. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:1979–1993.

Takeru Miyato, Shin-Ichi Maeda, Masanori Koyama, and Shin Ishii. 2019. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1979–1993.

Ercong Nie, Shuzhou Yuan, Bolei Ma, Helmut Schmid, Michael Färber, Frauke Kreuter, and Hinrich Schuetze. 2025. Decomposed prompting: Probing multilingual linguistic structure knowledge in large language models. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 646–659, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Cédric Villani. 2013. Optimal transport: Old and new.

Rui Wang and Ricardo Henao. 2021. Unsupervised paraphrasing consistency training for low resource named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5308, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.

Wenya Wang and Sinno Jialin Pan. 2018. Transition-based adversarial network for cross-lingual aspect extraction. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*.

Xiaodong Wang, Junbao Zhuo, Shuhao Cui, Shuhui Wang, and Yuejian Fang. 2024a. Learning invariant representation with consistency and diversity for semi-supervised source hypothesis transfer. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5125–5129.

Yong Wang, Ningchuang Yang, Duoqian Miao, and Qiuyi Chen. 2024b. Aspect-guided multi-graph convolutional networks for aspect-based sentiment analysis. *DATA INTELLIGENCE*, 6(3):771–791.

Chengyan Wu, Bolei Ma, Yihong Liu, Zheyu Zhang, Ningyuan Deng, Yanshu Li, Baolan Chen, Yi Zhang, Yun Xue, and Barbara Plank. 2025a. M-ABSA: A multilingual dataset for aspect-based sentiment analysis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2530–2557, Suzhou, China. Association for Computational Linguistics.

Chengyan Wu, Bolei Ma, Zheyu Zhang, Ningyuan Deng, Yanqing He, and Yun Xue. 2025b. Evaluating zero-shot multilingual aspect-based sentiment analysis with large language models. *International Journal of Machine Learning and Cybernetics*, 16(10):8079–8101.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*, volume 33, pages 6256–6268. Curran Associates, Inc.

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.

Mao Zhang, Yongxin Zhu, Zhen Liu, Zhimin Bao, Yunfei Wu, Xing Sun, and Linli Xu. 2023. Span-level aspect-based sentiment analysis via table filling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9273–9284, Toronto, Canada. Association for Computational Linguistics.

Wenxuan Zhang, Ruidan He, Haiyun Peng, Lidong Bing, and Wai Lam. 2021a. Cross-lingual aspect-based sentiment analysis with aspect term code-switching. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9220–9230. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510, Online. Association for Computational Linguistics.

Ran Zhou, Xin Li, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2022. ConNER: Consistency training for cross-lingual named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8438–8449, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2013. Collective opinion target extraction in Chinese microblogs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1840–1850, Seattle, Washington, USA. Association for Computational Linguistics.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2015. Clopinionminer: Opinion target extraction in a cross-language scenario. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(4):619–630.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016. 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.