Curriculum Knowledge Distillation for Emoji-supervised Cross-lingual Sentiment Analysis

Jianyang Zhang[1,2], **Tao Liang**[2], **Mingyang Wan**[2], **Guowu Yang**[1] and **Fengmao Lv**[3,✉]
[1]University of Electronic Science and Technology of China
[2]Bytedance [3]Southwest Jiaotong University
jianyangzhang@std.uestc.edu.cn; taoliangdpg@126.com;
wanmingyang@bytedance.com; guowu@uestc.edu.cn; fengmaolv@126.com

Abstract

Existing sentiment analysis models have achieved great advances with the help of sufficient sentiment annotations. Unfortunately, many languages do not have sufficient sentiment corpus. To this end, recent studies have proposed cross-lingual sentiment analysis to transfer sentiment analysis models from resource-rich languages to low-resource languages. However, these studies either rely on external cross-lingual supervision (e.g., parallel corpora and translation model), or are limited by the cross-lingual gaps. In this work, based on the intuitive assumption that the relationships between emojis and sentiments are consistent across different languages, we investigate transferring sentiment knowledge across languages with the help of emojis. To this end, we propose a novel cross-lingual sentiment analysis approach dubbed Curriculum Knowledge Distiller (CKD). The core idea of CKD is to use emojis to bridge the source and target languages. Note that, compared with texts, emojis are more transferable, but cannot reveal the precise sentiment. Thus, we distill multiple Intermediate Sentiment Classifiers (ISC) on source language corpus with emojis to get ISCs with different attention weights of texts. To transfer them into the target language, we distill ISCs into the Target Language Sentiment Classifier (TSC) following the curriculum learning mechanism. In this way, TSC can learn delicate sentiment knowledge, meanwhile, avoid being affected by cross-lingual gaps. Experimental results on five cross-lingual benchmarks clearly verify the effectiveness of our approach.

1 Introduction

Nowadays, sentiment analysis approaches perform very well by fine-tuning large-scale pre-trained language models (Yang et al., 2019; Xie et al., 2020; Wang et al., 2021). However, their success heavily relies on manual sentiment annotations. Therefore, they fail to recognize the sentiment polarities in low-resource languages that do not have sentiment supervision.

Still, recent studies have attempted to identify sentiment in unlabeled languages by transferring sentiment knowledge from resource-rich languages, (e.g., English), to low-resource languages (e.g., Japanese and Arabic) and propose the Cross-Lingual Sentiment Analysis (CLSA) task (Zhou et al., 2014). Generally, these CLSA approaches can be divided into two patterns, parallel corpus based approaches, and unsupervised approaches. Through cross-lingual supervision provided by parallel corpus, the former kind of approaches can align the semantic gaps between source language and target language by Auto-Encoder (Zhou et al., 2014), Neural Translation Model (Eriguchi et al., 2018) and Bilingual Word Embedding (Barnes et al., 2018). However, the availability of parallel corpus limits the usability of these works.

To avoid relying on external cross-lingual supervision, recent studies have paid attention to a more challenging setting dubbed Unsupervised Cross-Lingual Sentiment Analysis (UCLSA) (Feng and Wan, 2019; Fei and Li, 2020; Zhang et al., 2021). Under this setting, sentiment classifiers need to adapt to new languages without the help of parallel corpus. Cross-lingual language models such as m-BERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020) and LaBSE (Feng et al., 2020) have been prevalent in UCLSA. These models are pre-trained by multilingual corpus and then fine-tuned by source language sentiment supervision. In addition, Unsupervised Machine Translation Fei and Li (2020) and Multilingual Language Model Feng and Wan (2019) have been proposed to align source and target languages in UCLSA.

Although the above unsupervised approaches have calibrated semantic representation over source and target languages, their performances are still

✉ Corresponding author: F. Lv.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 864–875
December 7-11, 2022 ©2022 Association for Computational Linguistics

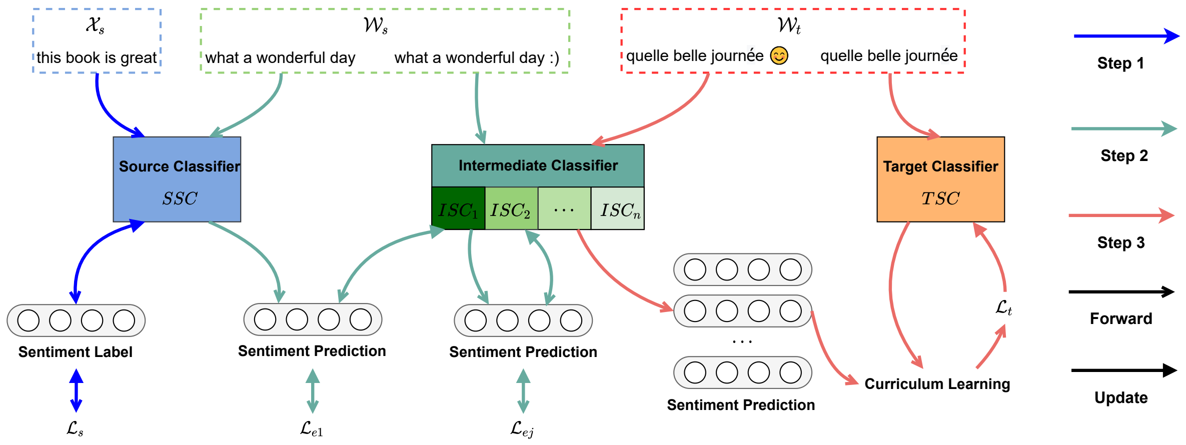


Figure 1: The overall architecture of our proposed CKD. CKD includes three kinds of sentiment classifiers, which are the **S**ource Language **S**entiment **C**lassifier SSC, the series of **I**ntermediate **S**entiment **C**lassifiers $\{ISC_1, ISC_2, ..., ISC_N\}$, and the **T**arget Language **S**entiment **C**lassifier TSC. In Step 1, we train SSC on $\mathcal{X}_s$. In Step 2, $\{ISC_1, ISC_2, ..., ISC_N\}$ are distilled from the former model of each of them to obtain models with different transferability and sentiment veracity on $\mathcal{W}_s$. ISC_1 is distilled from S and other ISC_j is distilled from ISC_{j-1}. Note that the attention weight of texts is decreased progressively with the model id j increased. Thus, the model with smaller attention weights on texts has more transferability but less precise sentiment knowledge. In Step 3, we distill $\{ISC_1, ISC_2, ..., ISC_N\}$ into TSC following the curriculum learning mechanism on $\mathcal{W}_t$. The losses of those more transferable ISCs begin with larger weights but then smaller. On the contrary, the weights of those less transferable ISCs begin with larger but then smaller.

unsatisfactory due to the large language discrepancy (Chen et al., 2017). In this work, we propose to use emojis to promote cross-lingual sentiment knowledge transfer, which can be dubbed Emoji-Supervised Cross-Lingual Sentiment Analysis (ESCLSA). ESCLSA is motivated by the intuitive assumption verified by Choudhary et al. (2018) that **the relationships between emojis and sentiments are consistent across different languages**. Thus, we can use multilingual corpus with emojis to bridge the cross-lingual gaps without manual cross-lingual supervision by aligning sentiment polarities with emojis. Notably, multilingual corpora with emojis is easy to obtain as it can be crawled from public social networks.

The most intuitive way to learn ESCLSA is manually labeling polarities of emojis, and using the emoji label as the pseudo-label of target language sentences. However, emojis fail to represent neutral sentiment or distinguish delicate polarities, e.g., very positive and positive. Meanwhile, the pseudo-label based method ignores the sentiment information in texts. Thus, it can not be well adapted to practical application. To handle these challenges, an end-to-end framework of ESCLSA needs to be established. ELSA (Chen et al., 2019) is the only previous approach for ESCLSA. It uses emoji prediction task to learn sentence representations. After that, ELSA uses these representations to classify sentiment polarities. The emoji-oriented representations are transferable. However, they ignore the information carried in text, which limits the performance of ELSA.

In this paper, we propose a novel ESCLSA approach dubbed Curriculum Knowledge Distiller (CKD). The overall architecture of CKD is shown in Figure 1. Texts can reveal the precise sentiment but are hard to transfer due to the cross-lingual gaps. On the other hand, emojis are more transferable, but the sentiment information in texts will be underestimated if we focus too much on the emojis (Chen et al., 2019). To address this limitation in the previous work Chen et al. (2019), we propose to distill multiple **I**ntermediate **S**entiment **C**lassifiers (ISCs) from **S**ource Language **S**entiment **C**lassifier (SSC) on source language tweets with emojis. By adjusting the attention weights of texts in ISCs, ISCs has different transferability and sentiment veracity. The models which are trained with low attention weights of texts (i.e., emoji-dominant ISCs), are more transferable, while the models with high attention weights of texts (i.e., text-dominant ISCs) pay more attention to extract sentiment information from texts. Based on these ISCs, we distill the **T**arget Language **S**entiment **C**lassifier (TSC) following the curriculum learning mechanism (Bengio et al., 2009). The TSC is trained by distillation losses computed from different intermediate models. Specifically, in the beginning of distilling TSC, the weights of emoji-dominant ISCs are large, as they are easier to transfer. During the distillation process, we progressively increase the weights of text-dominant ISCs to learn more precise sentiment knowledge from texts. Compared with the previous works, CKD can integrate more transferable information from emojis with more precise sentiment information in texts. Thus, CKD achieves the balance between transferbility and sentiment veracity. We verify our approach on five cross-language sentiment analysis benchmarks. The experimental results clearly support the effectiveness of our approach.

To sum up, the contributions of this work are three-fold:

- We propose a novel cross-lingual sentiment classification approach CKD which avoids the language discrepancy problem with the help of corpus with emojis in multilingual public social networks.
- We propose to distill multiple ISCs with different attention weights of texts to balance the transferability and the sentiment veracity. Furthermore, we propose to use the curriculum learning mechanism to distill these ISCs into the target language sentiment classifier. In this way, our model can transfer delicate sentiment knowledge across cross-lingual gaps.
- We conduct extensive experiments on five language pairs involving 11 tasks to evaluate our approach. CKD outperforms all the baseline models. That is, we provide a practical approach for cross-lingual sentiment analysis without requiring cross-lingual supervision.

2 Related Work

This section briefly reviews works related to ours, including cross-lingual sentiment analysis, knowledge distillation, and curriculum learning.

2.1 Cross-Lingual Sentiment Analysis

CLSA aims at transferring source language trained sentiment models to adapt target language (Zhou et al., 2014; Eriguchi et al., 2018; Barnes et al., 2018; Fei and Li, 2020). The early works of CLSA rely on cross-lingual supervision. Hajmohammadi et al. (2014); Al-Shabi et al. (2017); Chen et al. (2019) use Neural Translation Model to align source and target languages. In addition, Bilingual Word Embedding (Ziser and Reichart, 2018) and parallel corpus (Xu and Yang, 2017) are also used to bridge the cross-lingual gaps. Recent studies propose the unsupervised CLSA (UCLSA) setting to avoid relying on the expensive parallel corpus. Adversarial Training (Chen et al., 2018), Cross-Lingual Language Model (Feng and Wan, 2019) and Unsupervised Machine Translation (Fei and Li, 2020) have been introduced to achieve CLSA without any cross-lingual supervision. Still, these UCLSA approaches are limited when the gaps between source and target languages are very large, e.g., English to Japanese (Chen et al., 2019). Thus, we propose to use emojis to promote cross-lingual sentiment knowledge transfer. Emojis represent consistent sentiments across different languages (Choudhary et al., 2018), which can be the bridge across cross-lingual gaps.

2.2 Emoji Based Sentiment Analysis

Existing works have noticed the relevance between emojis and sentiment. Liu et al. (2021) directly inputs emojis as the following words into the sentiment classifier to improve sentiment analysis, while Lou et al. (2020) proposes to fuse emoji sentiment into text embedding by the attention algorithm and Yuan et al. (2021) proposes to improve emoji embedding by the graph network. In addition, Chaudhary et al. (2019) uses emojis for irony detection. Although these studies have explored the role of emojis in sentiment analysis, these works are designed for the monolingual setting and require manual annotations on texts with emojis. Thus, these works cannot promote CLSA with the help of emojis.

ELSA (Chen et al., 2019) is the first work of ESCLSA, which uses the emoji prediction task to learn sentence representations. After that, it uses these representations to classify sentiment polarities. The emoji-oriented representations are transferable. However, it ignores the information carried in texts, which limits the performance of ELSA. In this paper, we propose a novel Emoji-supervised CLSA approach dubbed Emoji Knowledge Distillation to achieve the balance between transferability and sentiment veracity.

2.3 Knowledge Distillation

Knowledge distillation transfers knowledge from teacher model to student model by prompting student model to learn the soft output of teacher model (Hinton et al., 2015). In recent years, knowledge distillation techniques have been wildly used in cross-lingual studies. These studies use supervision in source language to train the teacher model and distill it into unlabeled target language to calibrate the domain gaps for Cross-Lingual Conversation (Sun et al., 2021), Cross-Lingual Sentiment Analysis (Xu and Yang, 2017), Cross-Lingual Semantic Relation Classification (Vyas and Carpuat, 2019), and Cross-Lingual Named Entity Recognition (Wu et al., 2020; Li et al., 2022). Note that Xu and Yang (2017) is the previous work introducing distillation in CLSA. However, it distills the target language classifier on parallel corpora, which is not available in our setting. Following these works, we further propose to distill multiple intermediate models from source language sentiment classifier to obtain both high transferability and high sentiment veracity models.

2.4 Curriculum Learning

Curriculum learning aims at solving the challenge of "how to study", by referring to the human learning strategy of studying from the easy samples to the hard samples (Bengio et al., 2009). The key challenge of curriculum learning is finding the rank of samples from easy to hard, as well as determining the timing of introducing hard samples (Soviany et al., 2022). Recent studies determine the learning sequence by the sentence length and the word rarity (Platanios et al., 2019), the learning curve of model (Matiisen et al., 2019), and Meta Learning (Zhan et al., 2021). In our approach, the transferability of intermediate models increases with the attention weight of texts decreasing. Therefore, we distill ISCs into TSC in the order of easy-to-transfer to hard-to-transfer, following the curriculum learning mechanism.

Algorithm 1 CKD Training Pipeline

```
$SSC^m, ISC_j^m, TSC^m \leftarrow$ SSC, the j-th ISC, and TSC models at iteration $m$;
$SSC^{m-1}, ISC_j^{m-1}, TSC^{m-1} \leftarrow$ SSC, the j-th ISC, and TSC models at iteration $m-1$;
$\mathcal{X}_s \leftarrow$ source language labeled samples;
$\mathcal{W}_s \leftarrow$ source language unlabeled tweets with emojis;
$\mathcal{W}_t \leftarrow$ target language unlabeled tweets with emojis;
# Train SSC
for $m = 1, 2, ..., M$ do
    $x_s, y_s \leftarrow \mathcal{X}_s^m$;
    $SSC^m \leftarrow$ Adam model update $\nabla \mathcal{L}_s\left(x_s, y_s\right)$;

    $SSC^{m-1} \leftarrow SSC^m$;
end for
# Distill $\{ISC_1, ISC_2, ..., ISC_n\}$
for $m = 1, 2, ..., M$ do
    $w_s, w_s \leftarrow \mathcal{W}_s^m$;
    $ISC_1^m \leftarrow$ Adam update $\nabla \mathcal{L}_{e1}\left(w_s, w_s^*\right)$;
    $ISC_1^{m-1} \leftarrow ISC_1^m$;
end for
for $j = 2, ..., n$ do
    for $m = 1, 2, ..., M$ do
        $w_s \leftarrow \mathcal{W}_s^m$;
        $ISC_j^m \leftarrow$ Adam update $\nabla \mathcal{L}_{ej}\left(w_s\right)$;
        $ISC_j^{m-1} \leftarrow ISC_j^m$;
    end for
end for
# Distill $TSC$
for $m = 1, 2, ..., M$ do
    $w_t, w_t^* \leftarrow \mathcal{W}_t^m$;
    for $j = 1, ..., n$ do
        $\beta_j^m \leftarrow \frac{1+\lambda_j \cos\left(\frac{m\gamma}{M}\pi\right)}{2}$;
    end for
    $TSC^m \leftarrow$ Adam update $\nabla \mathcal{L}_t^m\left(w_t, w_t^*\right)$;
    $TSC^{m-1} \leftarrow TSC^m$;
end for
```

3 Methodology

In this section, we introduce the methodology of our proposed CKD. The key idea of our approach is using emojis that express the consistent sentiment in source and target languages to distill sentiment knowledge across languages. Therefore, we use source language supervision to train SSC, firstly. Then, as emojis can not explain such precise sentiment as texts but are easier to transfer, we propose to distill multiple ISCs with various attention weights of texts. The sentiment knowledge learned by ISCs with higher text attention weights is more precise but less transferable. Finally, we distill ISCs into TSC following the curriculum learning mechanism. Specifically, for the ISCs that are easy to transfer, we assign high weights of their distillation losses first, then decrease the weights during training. On the contrary, for the ISCs that are hard to transfer, we reverse the direction of weight change. In this way, TSC can learn precise sentiment knowledge from those hard-to-transfer ISCs, meanwhile avoiding the language discrepancy problem with the help of those easy-to-transfer ISCs.

In the following subsections, we will introduce the implementation details of our proposed approach. The overall architecture of CKD is shown in Figure 1.

3.1 Problem Description

In ESCLSA, we are given the labeled source language samples without emojis $\mathcal{X}_s = \{(x_s, y_s)\}$, where $y_s \in \{1, ..., c\}$, c is the number of sentiment polarities. The unlabeled source/target language tweets with emojis $\mathcal{W}_s = (w_s, w_s^*) / \mathcal{W}_t = (w_t, w_t^*)$, where w_s and w_s^* represent a source language tweet retained and eliminated emojis, respectively.

3.2 Distillation in Source Language

In this section, we learn several intermediate classifiers $\{ISC_1, ISC_2, ..., ISC_n\}$ with different attentions on the emojis to consider both sentiment veracity and cross-lingual transferability. We train a source language sentiment classifier SSC by $\mathcal{X}_s$ to distill source language sentiment knowledge, firstly, formally,

$$\mathcal{L}_s(x_s, y_s) = -\sum_{i=1}^{c} I_{i=y_s} \cdot \log(SSC(x_s)), \tag{1}$$

where $\mathcal{L}_s$ is a cross-entropy loss function, and $I_{i=y_s}$ is an indication function.

As discussed in Section 3, we distill SSC into multiple intermediate models $\{ISC_1, ISC_2, ..., ISC_n\}$, where n is the number of ISCs. For a latter model ISC_j, we set the attention weights of words to be smaller. Therefore, with the increase of j, model ISC_j pays more attention to emojis, the veracity of sentiment is lower, and the transferability is higher. We use the Transformer (Vaswani et al., 2017) like architecture for ISCs. Thus, the attention weights of words can be adjusted by modifying Eq. (1) in Vaswani et al. (2017) to

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}} - \alpha_j E\right) V, \tag{2}$$

where α_j is the hyper-parameter that represents the attenuation rate of words for M_j, and $E \in \{0, 1\}^k$ is the word mask vector, in which 1 means the current token is a word, on the country 0 means it is an emoji, k is the sentence length. Using adjusted attention function Eq. 2, we distill ISC_1 from SSC by cross-entropy loss, formally,

$$\mathcal{L}_{e1}(w_s, w_s^*) = -\sum_{i=1}^{c} SSC^i(w_s^*) \log\left(ISC_1^i(w_s)\right). \tag{3}$$

Note that as SSC is trained on $\mathcal{X}_s$, which is a text dataset without emojis, emojis are eliminated for the input of SSC. Then we distill the rest intermediate models $ISC_2, ..., ISC_n$ from the previous model to gradually improve the transferability, formally,

$$\mathcal{L}_{ej}(w_s) = -\sum_{i=1}^{c} ISC_{j-1}^i(w_s) \log\left(ISC_j^i(w_s)\right), \tag{4}$$

where $2 \leq j \leq n$. Through the above steps, a series of intermediate models with different transferability and sentiment veracity is established.

3.3 Curriculum Distillation in Target Language

We use curriculum learning to integrate the above classifiers $\{ISC_1, ISC_2, ..., ISC_n\}$ into the target language sentiment classifier TSC to infer the sentiment polarity in the target language. Inspired by the "from easy to hard" learning strategy of curriculum learning (Bengio et al., 2009), we trained TSC by distillation losses computed from different intermediate models. Specifically, in the beginning of distilling TSC, the weights of emoji-dominant ISCs are large. During the distillation process, we progressively increase the weights of text-dominant ISCs. To achieve this goal, the weight of each ISC changes smoothly in the iteration by an adjustable weight function, formally,

$$\beta_j^m = \frac{1 + \lambda_j \cos\left(\frac{m\gamma}{M}\pi\right)}{2}, \tag{5}$$

where $\lambda_j \in [-1, 1]$ is the hyper-parameter to determine the slope, m is the current iteration number, γ is the hyper-parameter to determine the stop state, and M is the total iteration number. The distillation loss function $\mathcal{L}_t^{(j)}$ of each ISC_j in iteration m is shown as,

$$\mathcal{L}_t^{(j)}(w_t, w_t^*) = -\sum_{i=1}^{c} ISC_j^i(w_t^m) \log\left(TSC^i(w_t^{m*})\right). \tag{6}$$

Thus, the overall objective function $\mathcal{L}_t^m$ can be formulated as,

$$\mathcal{L}_t^m(w_t, w_t^*) = \frac{1}{\sum\limits_{j=1}^{n} \beta_j^m} \sum_{j=1}^{n} \left(\beta_j^m \mathcal{L}_t^{(j)}\right). \tag{7}$$

Dataset	Language	Number of Classes	Number of Samples
Amazon Review (Duh et al., 2011)	Multi	2	4,000
Yelp (Zhang et al., 2015)	English	5	700,000
Hotel Review (Lin et al., 2015)	Chinese	5	20,000
Social Media Posts (Mohammad et al., 2016)	Arabic	3	1,000

Table 1: Statistics for Amazon Review (Duh et al., 2011), Yelp (Zhang et al., 2015), Hotel Review (Lin et al., 2015), and Social Media Posts (Mohammad et al., 2016) datasets in terms of Language, number of classes, and number of samples. Note that as Amazon Review is a multilingual dataset, we show the number of samples for each task in each language.

Note that, in the consideration of insuring T to learn sentiment knowledge in the target language, we use the tweets eliminated emojis w_t^* as the input of T.

3.4 Training

In this section, we introduce the training strategy for CKD. First, we use source language supervised dataset $\mathcal{X}_s$ to train the source language sentiment classifier SSC. Then we distill SSC to multiple intermediate models $\{ISC_1, ISC_2, ..., ISC_m\}$ where the text attention weights of these models are from high to low. Moreover, we set the curriculum schedule in the reverse order, i.e., from ISC_m to ISC_1. Finally, we distill the target language sentiment classifier T according to this schedule by the adjustable weight function β_j^m. Algorithm 1 describes our iterative training pipeline in detail. More training details are described in Section 4.2.

4 Experiments

In this section, we conduct extensive experiments to verify the effectiveness of our proposed approach. Following Fei and Li (2020), eleven cross-lingual tasks based on five language pairs are used in the experiments. All those pairs take English to be the source language, and the target languages are German, French, Japanese, Chinese and Arabic, respectively.

4.1 Datasets

To leverage the supervision of emojis, we establish a Twitter corpora with emojis for each source and target language pair. We collect 20,000 unlabeled tweets with emojis for each language posted in December 2018 and use sentiment emojis listed by Yin et al. (2021).

For labeled data, as shown in Table 1, we employ samples in six different languages (i.e., English, German, French, Japanese, Chinese and Arabic) from the following datasets:

Amazon Review (Duh et al., 2011): This dataset consists of four languages (i.e., English, German, French and Japanese) for the binary sentiment classification problem, while each language contains three domains (i.e., Books, DVD, and Music). For each cross-lingual task, there are 2,000 samples for train and 2,000 for test.

Yelp (Zhang et al., 2015): It is a large-scale English sentiment dataset containing 700K reviews from five classes. We use the original class tags of Yelp for the English-Chinese pair, and convert it into 3 sentimental levels, i.e., $1, 2$ →"negative", 3 →"neutral", and $4, 5$ →"positive", for the English-Arabic pair.

Hotel Review (Lin et al., 2015): This dataset contains 170K Chinese hotel reviews from 5 classes as in the Yelp dataset, and it is used for the English-Chinese pair in the experiments. Following Fei and Li (2020) we randomly sample 20K reviews for test.

Social Media Posts (Mohammad et al., 2016): It is an Arabic sentiment dataset with three sentimental labels(i.e., positive, neutral and negative). We randomly sample 1000 instances and use them for testing on the English-Arabic pair.

4.2 Experimental Setup

We compare our proposed CKD with a number of baseline methods under different categories: i) methods with explicit cross-lingual supervision (i.e., LR+MT, CR-RL (Xiao and Guo, 2013), Bi-PV (Pham et al., 2015) and CLDFA (Xu and Yang, 2017)); ii) methods with implicit cross-lingual supervision (i.e., UMM (Xu and Wan, 2017), PBLM (Ziser and Reichart, 2018), DAN (Chen et al., 2018), mSDA (Chen et al., 2012) and ADAN (Chen et al., 2018)); iii) methods with no cross-lingual supervision (i.e., BWE (Conneau

	Approach	German (2)				French (2)				Japanese (2)			
		Books	DVD	Music	Avg	Books	DVD	Music	Avg	Books	DVD	Music	Avg
$\mathcal{S}$	LR+MT	79.68	77.92	77.22	78.27	80.76	78.83	75.78	78.46	70.22	71.30	72.02	71.18
	CR-RL	79.89	77.14	77.27	78.10	78.25	74.83	78.71	77.26	71.11	73.12	74.38	72.87
	Bi-PV	79.51	78.60	82.45	80.19	84.25	79.60	80.09	81.31	71.75	75.40	75.45	74.20
	CLDFA	83.95	83.14	79.02	82.04	83.37	82.56	83.31	83.08	77.36	80.52	76.46	78.11
$\mathcal{I}$	UMM	81.65	81.27	81.32	81.41	80.27	80.27	79.41	79.98	71.23	72.55	75.38	73.05
	PBLM	78.65	79.90	80.10	79.50	77.90	75.65	75.95	76.50	-	-	-	-
$\mathcal{U}$	BWE	76.00	76.30	73.50	75.27	77.80	78.60	78.20	78.17	55.93	57.55	54.35	55.94
	MAN-MoE	82.40	78.80	77.15	79.45	81.10	84.25	80.90	82.08	62.78	69.10	72.60	68.16
	m-BERT	84.35	82.85	83.85	83.68	84.55	85.85	83.65	84.68	73.35	74.80	76.10	74.75
	XLM-R	89.05	86.40	87.15	87.53	88.35	87.88	83.55	86.59	82.40	83.95	82.85	83.07
	CLIDSA	86.65	84.60	85.05	85.43	87.20	87.95	87.15	87.43	79.35	81.90	84.05	81.77
	MVEC	88.41	87.32	89.97	88.61	89.08	88.28	88.50	88.62	79.15	77.15	79.70	78.67
$\mathcal{E}$	ELSA	86.40	86.10	87.80	86.77	86.00	85.70	86.00	85.90	78.30	79.10	80.80	79.40
	CKD	**92.45**	**90.15**	**91.55**	**91.38**	**91.95**	**91.35**	**89.65**	**90.98**	**84.10**	**85.30**	**86.20**	**85.20**

Table 2: Comparisons with the set baselines over Amazon Review Dataset (Duh et al., 2011). $\mathcal{S}$ represents the models are with explicit cross-lingual supervision, $\mathcal{I}$ represents the models are with implicit cross-lingual supervision, $\mathcal{U}$ represents the models are without cross-lingual supervision, and $\mathcal{E}$ represents the emoji supervised cross-lingual sentiment analysis model. The highest performance is in bold.

Approach	Chinese (5)	Arabic (3)
LR+MT	34.01	51.67
DAN	29.11	48.00
mSDA	31.44	48.33
ADAN	42.49	52.54
m-BERT	38.85	50.40
XLM-R	46.60	51.16
MVEC	43.36	49.70
CKD	**49.86**	**53.14**

Table 3: Comparisons with the set baselines over English-Chinese (Zhang et al., 2015; Lin et al., 2015) and English-Arabic (Zhang et al., 2015; Mohammad et al., 2016). The highest performance is in bold.

	German	French	Japanese
CKD	**91.38**	**90.98**	**85.20**
w/o curriculum learning	89.98	89.53	84.05
w/o attenuation of text attention	89.30	88.47	83.80
source model	87.53	86.59	83.07

Table 4: Ablation study for the contribution of each design over the Amazon Review dataset. The performances are the average results for the three domains, Books, DVD, and Music.

et al., 2017), MAN-MoE (Chen and Qian, 2019), m-BERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), CLIDSA (Feng and Wan, 2019), MVEC (Fei and Li, 2020)); iv) the emoji supervised cross-lingual sentiment analysis model (i.e., ELSA (Chen et al., 2019)). It is worth noting that our CKD belongs to the fourth category, which makes use of emojis to bridge different languages. In addition, the compared baselines are validated on different benchmarks. Hence, we use different baselines in Table 2 and Table 3 for fairness.

In our proposed CKD, we use XLM-R (Conneau et al., 2020) to be the backbone for all SSC, ISCs, and TSC. XLM-R is pre-trained on CommonCrawl (Wenzek et al., 2020) which is a large-scale unsupervised multilingual corpus. All the CKD components are optimized by Adam optimizer (Kingma and Ba, 2014) with the learning rate of 3×10^{-6} and the mini-batch size of 28. The iteration time M for each step is $4,000$, and the early stop rate γ is 0.8. In addition, we set the number of ISCs to be 4, to balance the training costs and the performance of our approach. For these ISCs, the attenuation rate of words α are $(0, 0.1, 0.5, 1)$, and the slope parameters λ are $(-1, -0.5, 0.5, 1)$. Moreover, we use class average accuracy to evaluate the performance of CKD and baselines.

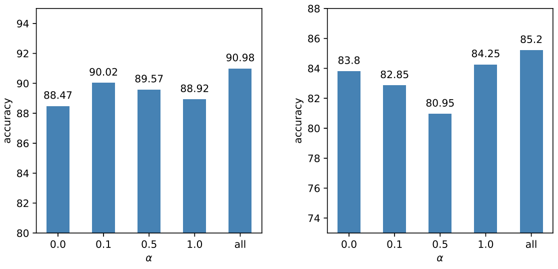
4.3 Comparison

The classification results on Amazon Review Dataset, English-Chinese, and English-Arabic pairs are shown in Table 2 and Table 3 respectively (where (n) indicates the number of sentiment polarity). Our proposed method outperforms all the compared methods in all target languages. Specifically, as shown in Table 2, our model significantly improves binary sentiment classification performance in German, French, and Japanese, achieving 2.36% to 3.43% improvements on each language averagely. As for multi-class sentiment classification, our approach achieves 0.6% to 3.26% improvements in Chinese and Arabic respectively. Note that, Japanese, Chinese, and Arabic are further from the source language, compared with German and French. The improvements of CKD are greater in these languages. This shows that emojis is a powerful bridge to transfer sentiment knowledge across languages, especially when the cross-lingual gaps are large.

In addition, compared with the early emoji powered work ELSA (Chen et al., 2019), our approach also brings significant improvements. The reason for this is that our model distills sentiment knowledge from source language and emojis, rather than using the emoji prediction task to learn cross-lingual consistent representations like ELSA does. The sentence representations learning from emoji prediction task only involves knowledge correlation to emojis. On the contrary, our TSC learns more sentiment knowledge from source language with the help of those ISCs with high attention weights to texts.

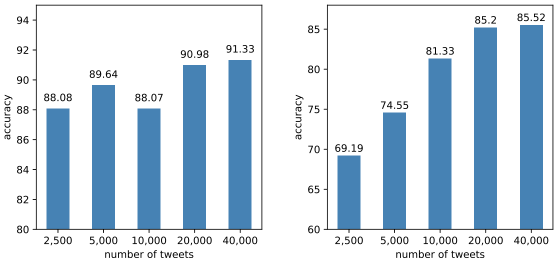
As for the baseline models, we can see that language model pre-trained approaches, i.e., m-BERT and XLM-R achieve impressive performances, which shows the effectiveness of pretraining on large-scale unsupervised multilingual corpus and the high semantic comprehension abilities of Transformer (Vaswani et al., 2017) based architectures. On the other hand, ELSA also performs well. ELSA uses the emoji prediction task to generate sentence embedding for both source and target languages. The high cross-lingual consistency of emojis brings ELSA the excellent ability of cross-lingual sentiment transfer. Still, they ignore the information carried in texts. Thus, we can achieve better performance than ELSA as discussed above.

4.4 Analysis



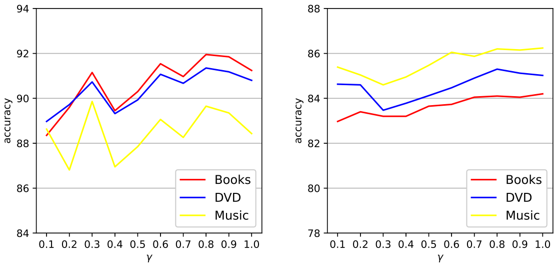
(a) French (b) Japanese

Figure 2: Average accuracy of TSCs distilled from ISCs with different attenuation of text attention weights α in (a) French and (b) Japanese. The last column "all" represents the result of curriculum learning.



(a) French (b) Japanese

Figure 3: Average accuracy of TSCs with different number of tweets for both source and target languages in (a) French and (b) Japanese.



(a) French (b) Japanese

Figure 4: Sensitivity analysis for the early stop rate γ from 0.1 to 1 in (a) French and (b) Japanese.

As shown in Table 4, we conduct the ablation study to demonstrate the contributions of each component in CKD. The first line represents the model trained with all proposed components. The next two lines represent the models without curriculum learning (Eq. (5)) and without attenuation of text attention (Eq. (2)), respectively. Note that as removing Eq. (2) makes ISCs no different, the third line represents the results of using the single ISC with attenuation rate $\alpha = 0$. In addition, the last line represents directly evaluating SSC on target language. It is clear that all the key parts of CKD generally make good contributions to promote cross-lingual sentiment knowledge transfer. Specifically, from the second line, the curriculum learning mechanism promotes TSC to learn ISCs flexibly, which achieves higher performances than averagely distilling ISCs. From the third line, we can see that even without attenuating the attention weight of texts, distillation still works. This is mainly caused by the soft labels in knowledge distillation that can filter samples which are hard to transfer. Another reason is that the model gains the cross-lingual transfer ability by paying attention to emojis.

In addition, we recommend distilling multiple ISCs with different attenuation of text attention weights. To evaluate the effectiveness of these ISCs, we test the performance of distilling TSC from each single ISC. Results are shown in Figure 2. TSC distilled from all ISCs performs better than those distilled from a single ISC. As for the performances of distilled from a single ISC, they show different patterns in different languages. In French, the ISC with 0.1 attenuation of texts performs best, but in Japanese, the ISC with the highest attenuation of texts performs best. The reason for this is that French is closer to English and shares more tokens in the tokenizer of XLM-R. On the contrary, Japanese is far from English, which gives emojis a more important role in cross-lingual sentiment transfer. Notably, in real applications, the test set of target language is unavailable. Thus, it is not feasible to manually select the attenuation of texts in ISCs.

As shown in Figure 3, we further analyze the effect of the size of unlabeled Twitter corpora on model performance. Generally, with a larger size of Twitter corpora, CKD can transfer more knowledge from the source language to the target language, and the performance will be better. 20,000 tweets for both source and target languages is the trade-off between the training cost and the performance. We can see that for Japanese, the performance decreases more severely with the decreasing of Twitter corpora size. This is mainly caused by the larger cross-lingual gap between English and Japanese compared with English-French pair. When the cross-lingual gap is large, more tweet samples are required for achieving cross-lingual knowledge transfer.

Moreover, we conduct the sensitivity analysis for the early stop rate γ to demonstrate the robustness of our model displayed in Figure 4. From this, we can see that the accuracy is sensitive to γ, as it determines the weight of Eq. 6 for each ISC at the end of distilling TSC. In the beginning, when TSC learns more from ISCs that are hard to transfer, the accuracy is higher. However, if γ is too large, the performance will degrade. This is because ISCs which pay much attention to texts will overfit the source language. Following the curriculum learning mechanism, TSC is first distilled from emoji-dominant ISCs , and then from text-dominant ISCs. Experimental results shown in Figure 4 verify the effectiveness of this strategy.

5 Conclusion

Emojis are widely used in social networks of various languages. Based on the intuitive assumption that the relationships between emojis and sentiments are cross-lingual consistent, we use the unsupervised multilingual corpus with emojis to transfer sentiment knowledge across languages. To achieve this goal, we propose CKD, a novel curriculum sentiment knowledge distillation framework. Using the source language supervision trained model, we distill a series ISCs. These ISCs have different transferability and sentiment veracity. Finally, we distill RSCs into TSC under the curriculum learning mechanism. That is, in the beginning of training, the weights of emoji-dominant ISCs are large, as they are easier to transfer. Then we increase the weights of text-dominant ISCs to learn more precise sentiment knowledge. In this way, CKD learns precise sentiment knowledge, meanwhile avoiding the language discrepancy problem. Extensive experiments on five languages involving 11 tasks verify the effectiveness of our approach.

Limitations

Our approach needs to distill ISCs one after another. Although the time complexity of such a distillation strategy used in our proposed CKD is a bit high, it boosts the classification performance by a considerable margin. Moreover, it is worth mentioning that the scope of this cross-lingual work focuses on pairs of languages. In the future, we will extend our CKD to learn one universal model that caters for all target languages at the same time.

Acknowledgments

This project is supported by the National Natural Science Foundation of China (No. 62106204 and 62172075), the Sichuan Natural Science Foundation (No. 2022NSFSC0911), the Chengdu Innovation and Technology Project (No. 2021-YF05-02414-GX) and the Fundamental Research Funds for Central Universities of China (No. 2682022CX068). J. Zhang and T. Liang contribute equally to this work.

References

Adel Al-Shabi, Aisah Adel, Nazlia Omar, and Tareq Al-Moslmi. 2017. Cross-lingual sentiment classification from english to arabic using machine translation. *International journal of advanced computer science and applications*, 8(12).

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.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Aditi Chaudhary, Shirley Anugrah Hayati, Naoki Otani, and Alan W Black. 2019. What a sunny day: toward emoji sensitive irony detection. *Proc. W-NUT*, page 212.

Minmin Chen, Zhixiang Xu, Kilian Weinberger, and Fei Sha. 2012. Marginalized denoising autoencoders for domain adaptation. *arXiv preprint arXiv:1206.4683*.

Qiang Chen, Chenliang Li, and Wenjie Li. 2017. Modeling language discrepancy for cross-lingual sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 117–126.

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.

Zhenpeng Chen, Sheng Shen, Ziniu Hu, Xuan Lu, Qiaozhu Mei, and Xuanzhe Liu. 2019. Emoji-powered representation learning for cross-lingual sentiment classification. In *The World Wide Web Conference*, pages 251–262.

Zhuang Chen and Tieyun Qian. 2019. Transfer capsule network for aspect level sentiment classification. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 547–556.

Nurendra Choudhary, Rajat Singh, Vijjini Anvesh Rao, and Manish Shrivastava. 2018. Twitter corpus of resource-scarce languages for sentiment analysis and multilingual emoji prediction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1570–1577.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard 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.

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

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.

Kevin Duh, Akinori Fujino, and Masaaki Nagata. 2011. Is machine translation ripe for cross-lingual sentiment classification? In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 429–433.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv preprint arXiv:1809.04686*.

Hongliang Fei and Ping Li. 2020. Cross-lingual unsupervised sentiment classification with multi-view transfer learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5759–5771.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

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.

Mohammad Sadegh Hajmohammadi, Roliana Ibrahim, and Ali Selamat. 2014. Bi-view semi-supervised active learning for cross-lingual sentiment classification. *Information Processing & Management*, 50(5):718–732.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Zhuoran Li, Chunming Hu, Xiaohui Guo, Junfan Chen, Wenyi Qin, and Richong Zhang. 2022. An unsupervised multiple-task and multiple-teacher model for cross-lingual named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 170–179.

Yiou Lin, Hang Lei, Jia Wu, and Xiaoyu Li. 2015. An empirical study on sentiment classification of chinese review using word embedding. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation: Posters*, pages 258–266.

Chuchu Liu, Fan Fang, Xu Lin, Tie Cai, Xu Tan, Jianguo Liu, and Xin Lu. 2021. Improving sentiment analysis accuracy with emoji embedding. *Journal of Safety Science and Resilience*, 2(4):246–252.

Yinxia Lou, Yue Zhang, Fei Li, Tao Qian, and Donghong Ji. 2020. Emoji-based sentiment analysis using attention networks. *ACM Transactions on asian and low-resource language information processing (TALLIP)*, 19(5):1–13.

Tambet Matiisen, Avital Oliver, Taco Cohen, and John Schulman. 2019. Teacher–student curriculum learning. *IEEE transactions on neural networks and learning systems*, 31(9):3732–3740.

Saif M Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130.

Hieu Pham, Minh-Thang Luong, and Christopher D Manning. 2015. Learning distributed representations for multilingual text sequences. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 88–94.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabás Póczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. 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 1162–1172.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum learning: A survey. *International Journal of Computer Vision*, pages 1–40.

Weiwei Sun, Chuan Meng, Qi Meng, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Conversations powered by cross-lingual knowledge. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1442–1451.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yogarshi Vyas and Marine Carpuat. 2019. Weakly supervised cross-lingual semantic relation classification via knowledge distillation. In *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Édouard Grave. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012.

Qianhui Wu, Zijia Lin, Börje F Karlsson, Jian-Guang Lou, and Biqing Huang. 2020. Single-/multi-source cross-lingual ner via teacher-student learning on unlabeled data in target language. *arXiv preprint arXiv:2004.12440*.

Min Xiao and Yuhong Guo. 2013. Semi-supervised representation learning for cross-lingual text classification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1465–1475.

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

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.

Ruochen Xu and Yiming Yang. 2017. Cross-lingual distillation for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Wenjie Yin, Rabab Alkhalifa, and Arkaitz Zubiaga. 2021. The emojification of sentiment on social media: Collection and analysis of a longitudinal twitter sentiment dataset. *arXiv preprint arXiv:2108.13898*.

Xiaowei Yuan, Jingyuan Hu, Xiaodan Zhang, and Honglei Lv. 2021. Pay attention to emoji: Feature fusion network with emograph2vec model for sentiment analysis.

Runzhe Zhan, Xuebo Liu, Derek F Wong, and Lidia S Chao. 2021. Meta-curriculum learning for domain adaptation in neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14310–14318.

Wenxuan Zhang, Ruidan He, Haiyun Peng, Lidong Bing, and Wai Lam. 2021. Cross-lingual aspect-based sentiment analysis with aspect term code-switching. In *Proceedings of the 2021 Conference on*

Empirical Methods in Natural Language Processing, pages 9220–9230.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Huiwei Zhou, Long Chen, and Degen Huang. 2014. Cross-lingual sentiment classification based on denoising autoencoder. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 181–192. Springer.

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.