SENTIX: A Sentiment-Aware Pre-Trained Model for Cross-Domain Sentiment Analysis

Jie Zhou[1,2], Junfeng Tian[3], Rui Wang[3], Yuanbin Wu[2*], Wenming Xiao[3], and Liang He[1,2*]

[1]Shanghai Key Laboratory of Multidimensional Information Processing
East China Normal University, China
[2]School of Computer Science and Technology
East China Normal University, China
jzhou@ica.stc.sh.cn, {ybwu,lhe}@cs.ecnu.edu.cn
[3]Alibaba Group, China
{tjf141457,masi.wr,wenming.xiaowm}@alibaba-inc.com

Abstract

Pre-trained language models have been widely applied to cross-domain NLP tasks like sentiment analysis, achieving state-of-the-art performance. However, due to the variety of users' emotional expressions across domains, fine-tuning the pre-trained models on the source domain tends to overfit, leading to inferior results on the target domain. In this paper, we pre-train a *sentiment-aware* language model (SENTIX) via domain-invariant sentiment knowledge from large-scale review datasets, and utilize it for cross-domain sentiment analysis task *without fine-tuning*. We propose several pre-training tasks based on existing lexicons and annotations at both token and sentence levels, such as emoticons, sentiment words, and ratings, without human interference. A series of experiments are conducted and the results indicate the great advantages of our model. We obtain new state-of-the-art results in all the cross-domain sentiment analysis tasks, and our proposed SENTIX can be trained with only 1% samples (18 samples) and it achieves better performance than BERT with 90% samples. Code is available at https://github.com/12190143/SentiX.

1 Introduction

Sentiment analysis has gained widespread attention from both industry and academia, which aims to judge the sentimental polarity of the given text (Liu, 2012). Most existing works heavily rely on labeled data to train separate sentiment classifiers for each domain, which are both expensive and time-consuming to obtain (Socher et al., 2013). Therefore, cross-domain sentiment analysis has become a promising direction, which transfers (invariant) sentiment knowledge from the source domain to the target domain[1].

The major challenge here is that language expressions for sentimental text usually vary across different domains. For instance, "fast" has a positive sentiment towards "service" in the restaurant domain (Figure 1), while in the laptop domain, "fast" expresses a negative sentiment for "power consumption". Furthermore, models trained on the source domain tend to overfit, since they learn domain-specific knowledge excessively. Therefore, many studies (Du et al., 2020; Ziser and Reichart, 2018; Li et al., 2018) propose to address this issue by extracting *domain-invariant* features.

Recently, pre-trained language models like BERT (Devlin et al., 2019) have achieved the state-of-the-art performance on multiple sentiment analysis tasks (Hoang et al., 2019; Munikar et al., 2019; Raffel et al., 2019). However, when they are directly applied to cross-domain sentiment analysis (Du et al., 2020), two problems arise: 1) Existing pre-trained models focus on learning the semantic content via self-supervision strategies, while ignoring *sentiment-specific* knowledge at the pre-training phrase; 2)

* Yuanbin Wu and Liang He are the corresponding authors of this paper. This work was conducted when Jie Zhou was interning at Alibaba DAMO Academy.

[1]Usually, we assume that there are abundant labeled data in the source domain, while little or no in the target domain. Thus, the model is trained on source domain and tested on the target domain for this task.

Proceedings of the 28th International Conference on Computational Linguistics, pages 568–579
Barcelona, Spain (Online), December 8-13, 2020

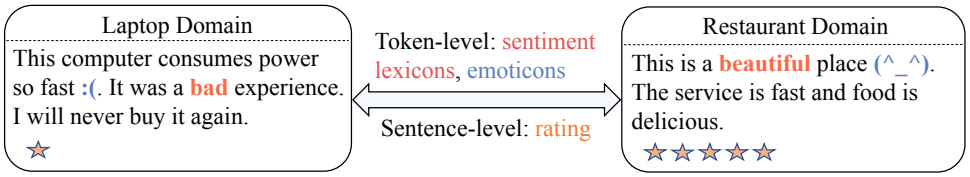


Figure 1: An example of sentiment knowledge (e.g., "bad", "beautiful", ":(", "(^_^)").

during the fine-tuning phase, pre-trained models may overfit the source domain by learning too much domain-specific sentiment knowledge, leading to degraded performance on the target domain.

To address the above-mentioned problems, we propose a *sentiment-aware* pre-trained model, named SentiX, to learn the *domain-invariant* sentiment knowledge at the pre-training phase, and it does *not* need to be fine-tuned for the cross-domain tasks. In particular, we observe that many widely available review datasets contain rich sentiment information, which can be utilized to enhance the domain-invariant knowledge acquisition. Large-scale review datasets, such as Yelp and Amazon, consist of 240 million reviews across 30 domains, full of sentiment words, emoticons and ratings. Taking Figure 1 as an example, these reviews contain opinion words (like "bad", "beautiful") and emoticons (like ":(", "(^_^)"), and their ratings are 1 and 5, respectively.

In order to obtain the above domain-invariant sentiment knowledge, we propose several sentiment-aware pre-training objectives, including token and sentiment prediction. At the token level, the sentiment words and emoticons are masked with a higher rate than the general words to emphasize the sentiment knowledge, and we pre-train SentiX to predict sentiment-aware words, emoticons, and token sentiments. At the sentence level, we introduce a rating prediction strategy to learn the sentiment knowledge based on the whole sentence.

We conduct extensive experiments on cross-domain sentiment analysis tasks to evaluate the effectiveness of SentiX, and obtain state-of-the-art results on all settings. SentiX achieves more than 90% accuracy over all cross-domain sentiment analysis datasets with only 1% samples, outperforming BERT trained with 90% samples. Through visualization of the feature representation, we observe that SentiX significantly reduces the overfitting issue, while the in-domain tests prove that our SentiX also obtains significant improvement over BERT for both the sentence-level and aspect-based sentiment classification tasks.

The main contributions of this paper can be summarized as follows:

- We propose SentiX for cross-domain sentiment classification to learn rich domain-invariant sentiment knowledge in large-scale unlabeled multi-domain data.
- We design several pre-training objectives at both token level and sentence level to learn such domain-invariant sentiment knowledge by masking and prediction.
- The experiments clearly show that SentiX obtains the state-of-the-art performance for cross-domain sentiment analysis and requires less annotated data than BERT to reach equivalent performances.

2 Preliminaries

Reviews contain a lot of semi-supervised sentiment signals, such as sentiment words, emoticons and ratings, and large-scale review data can be obtained from online review websites like Yelp. This sentiment knowledge can help learning domain-common sentiment feature for the cross-domain task.

- **Sentiment Words**. Sentiment lexicon contains a lot of sentiment information and is widely used in sentiment analysis. The words in lexicon are regarded as sentiment words. The words in positive and negative sentiment lexicons are labeled as "P" and "N" respectively. Words out of the lexicons are labeled as "0". HowNet[2] and opinion lexicon (Hu and Liu, 2004) are used as sentiment lexicons.

[2]http://www.keenage.com/html/c_index.html

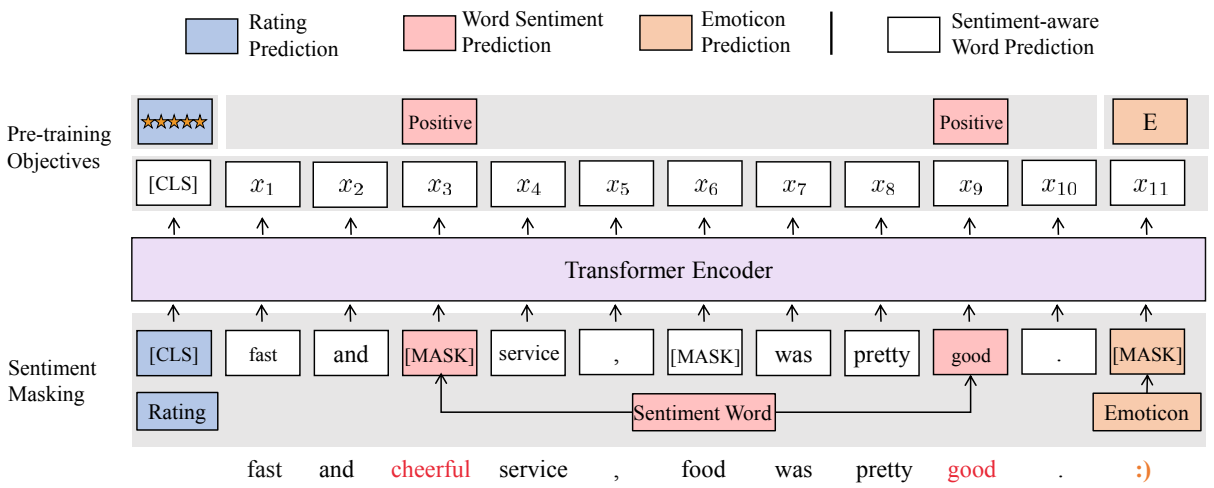


Figure 2: The framework of SENTIX. First, we design three sentiment masking strategies, including SWM (e.g., "cheerful", "good"), EM (e.g., ":)"), and GWM (e.g., "food"). Then, we propose four sentiment-aware prediction objectives from token level and sentence level.

- **Emoticons**. Emoticons are usually used by users in texts to express their emotion (Zhao et al., 2012). Each emoticon is made up of typographical symbols (e.g., ")", "(", ":", "D", "-") and denotes facial expressions. It can be read either sideways (e.g., a sad face ":-(") or normally (e.g.,a happy face "(^_^)")(Hogenboom et al., 2013). We extract the emoticons via regular expression and keep the top-100 emoticons in corpus (Table 1). If the words are matched by regular expression, they are labeled as "E", otherwise "0". [3]

- **Rating**. In addition to the above token-level sentiment knowledge, reviews contain sentence-level rating scores, which represent the overall sentiment polarities. Rating scores contain 5 level: very negative, negative, neutral, positive, and very positive. The scores' distribution is unbalanced, and we perform average sampling on the original data. Notably, labels of ratings are relatively difficult to be obtained than sentiment words and emoticons since only reviews data contain ratings. Therefore we study the ablation test in Section 5.3, which demonstrate that our model still performs better than the state-of-the-art approach without ratings.

To make full use of this rich sentiment knowledge for cross-domain sentiment analysis, we design several pre-training objectives to enhance the model with domain-invariant sentiment knowledge.

3 SENTIX

SENTIX is a sentiment-aware pre-training model for cross-domain sentiment analysis. It learns domain-invariant features from the above domain-invariant sentiment knowledge, including sentiment lexicons, emoticons, and ratings. The framework contains sentiment masking and pre-training objectives, as shown in Figure 2. Sentiment masking (Section 3.1) recognizes the sentiment information of an input sequence from sentiment knowledge. Pre-training objectives require encoder not only reconstruct the masked sentiment tokens, but also distinguish the word sentiment polarity, emoticon and rating (Section 3.2).

Formally, given a sentence $x = \{x_1, x_2, ..., x_{|x|}\}$, we first obtain a corrupted sentence $\hat{x} = \{\hat{x}_1, \hat{x}_2, ..., \hat{x}_{|\hat{x}|}\}$ ($\hat{x} \in \hat{\mathcal{X}}$, where $\hat{\mathcal{X}}$ is the corrupted corpus.) via sentiment masking. The sentiment-aware pre-training tasks are proposed to predict word x_i, sentiment s_i, emoticon e_i of in token level and rating r in sentence level. Here $s_i \in \{P, N, 0\}$ represents the sentiment polarity (positive, negative, others) of word x_i, $e_i \in \{E, 0\}$ indicates whether word x_i is an emoticon, and $r \in \{1, 2, 3, 4, 5\}$ is the rating of x.

[3]Emoticons do not occur in the vocabulary of BERT, thus we use unused tokens (e.g., [unused1], [unused2]) to represent them.

3.1 Sentiment Masking

Sentiment masking aims to enhance the sentiment information at the token level. Previous pre-trained models adopt masked language modeling (MLM) to learn semantic information. Some input tokens are randomly masked, and the goal is to predict these masked tokens. In addition to this general word masking, we propose sentiment word masking and emoticon masking for learning sentiment knowledge through recovering.

- **Sentiment Word Masking (`SWM`).** To enrich the sentiment information, we mask the sentiment words with 30% rate as these words are important for sentiment analysis[4].
- **Emoticon Masking (`EM`).** Since the number of emoticons in one sentence is relatively small and deleting emoticons will not influence the semantic information of the sentence, we mask 50% emoticons for each sentence.
- **General Word Masking (`GWM`).** If we only focus on the sentiment words and emoticons, SENTIX may lose the general semantic information of the other words. Thus, following the original BERT, we use [MASK] to replace the general word in sentence with 15% rate to learn the semantic information.

3.2 Pre-training Objectives

Sentiment masking produces corrupted sentences $\hat{x}$ where part of the sentiment words, emoticons and general words are substituted with masked tokens. Three token-level and one sentence-level prediction objectives are designed to learn the domain-invariant sentiment knowledge from the pre-training phase.

Sentiment-aware Word Prediction (`SWP`) Based on our sentiment masking strategies, the corrupted tokens that contain extra sentiment words and emoticons are obtained to capture the sentiment information. The corrupted sentence $\hat{x}$ is input to transformer encoder to obtain each word representations h_i and sentence representation $h_{[\text{CLS}]}$. Then a Softmax layer is used to compute each word's probability $P(x_i|\hat{x}_i) = \text{Softmax}(W_w \cdot h_i + b_w)$. The loss function L_w is the cross-entropy between the predicted probability and the true word label.

$$\mathcal{L}_w = -\frac{1}{|\hat{\mathcal{X}}|}\sum_{\hat{x}\in\hat{\mathcal{X}}}\frac{1}{|\hat{x}|}\sum_{i=1}^{|\hat{x}|}\log(P(x_i|\hat{x}_i))$$

Word Sentiment Prediction (`WSP`) According to the sentiment knowledge, we label the word's sentiment into positive, negative and others. Thus, we design `WSP` for learning the sentiment knowledge of the tokens. We aim to infer the sentiment polarity s_i of word x_i according to h_i, $P(s_i|\hat{x}_i) = \text{Softmax}(W_s \cdot h_i + b_s)$. The cross-entropy loss $\mathcal{L}_s = -\frac{1}{|\hat{\mathcal{X}}|}\sum_{\hat{x}\in\hat{\mathcal{X}}}\frac{1}{|\hat{x}|}\sum_{i=1}^{|\hat{x}|}\log(P(s_i|\hat{x}_i))$ is used here.

Emoticon Prediction (`EP`) To further capture the token-level sentiment knowledge, we propose to predict the emoticon label e_i of word x_i, $P(e_i|\hat{x}_i) = \text{Softmax}(W_e \cdot h_i + b_e)$. The loss $\mathcal{L}_e = -\frac{1}{|\hat{\mathcal{X}}|}\sum_{\hat{x}\in\hat{\mathcal{X}}}\frac{1}{|\hat{x}|}\sum_{i=1}^{|\hat{x}|}\log(P(e_i|\hat{x}_i))$ is also computed using cross-entropy function.

Rating Prediction (`RP`) Above tasks focus on learning the token-level sentiment knowledge. Ratings represent the overall sentiment score of the reviews in sentence level. Inferring the rating will bring in the sentence-level sentiment knowledge. Similar to BERT, we use the final state $h_{[\text{CLS}]}$ as the sentence representation. The rating is predicted by $P(r|\hat{x}) = \text{Softmax}(W_r \cdot h_{[CLS]} + b_r)$ and the loss is calculated based on the predicted rating distribution,

$$\mathcal{L}_r = -\frac{1}{|\hat{\mathcal{X}}|}\sum_{\hat{x}\in\hat{\mathcal{X}}}\log(P(r|\hat{x})). \tag{1}$$

[4]We obtain the best hyperparameters of sentiment masking by pre-training SENTIX on 20% of our pre-training dataset. More detailed analysis of these hyperparameters will be explored in future work.

#	Emoticon	Count	#	Emoticon	Count	#	Emoticon	Count	#	Emoticon	Count
1	:)	1,915,687	6	):	160,970	11	:/	67,615	16	(:	40,062
2	:(	438,339	7	:D	113,440	12	=)	66,156	17	:P	31,573
3	:-)	308,345	8	(8	86,546	13	:-(	64,391	18	;D	15,718
4	;)	259,724	9	<3	78,492	14	8:	53,073	19	:o)	12,952
5	);	178,795	10	;-)	74,247	15	8)	46,124	20	=(	11,917

Table 1: Statistics information of top-20 emoticons.

3.3 Joint Training

Finally, we jointly optimize the token-level objective $\mathcal{L}_T$ and the sentence-level objective $\mathcal{L}_S$. The overall loss is $\mathcal{L} = \mathcal{L}_T + \mathcal{L}_S$, where $\mathcal{L}_T = \mathcal{L}_w + \mathcal{L}_s + \mathcal{L}_e$ and $\mathcal{L}_S = \mathcal{L}_r$.

4 Experimental Setup

4.1 Datasets

Pre-training The pre-training phase is conducted on two large-scale datasets: Amazon review dataset (Ni et al., 2019) and Yelp 2020 challenge dataset[5]. Amazon dataset contains 233 million reviews within 29 domains (Ni et al., 2019). The total number of yelp reviews is about 8 million. We preprocess the text via NLTK[6] and transfer all the letters into lower. We filter the text that contains less than 50 tokens or more than 512 tokens and sample the rating data in dealing with class-imbalance problem. The statistics information of top-20 emoticons is shown in Table 1.

Sentiment Analysis To verify the effectiveness of SENTIX, We evaluate on the widely used cross-domain sentiment dataset (Blitzer et al., 2007), containing four domains: Books (B), DVD (D), Electronic (E), and Kitchen & Housewares (K). Following the setting of previous works (Ziser and Reichart, 2018; Qu et al., 2019), we test on 12 cross-domain tasks. The model is trained on the source domain and tested on the target domains. As before, we split 200 samples from 2000 samples in source domain as a development set to find the best hyperparameters. Besides, we examine the model on four popular sentient classification datasets: SST-2-Root (Socher et al., 2013), SST-5-Root (Socher et al., 2013), IMDB (Maas et al., 2011) and Yelp (Zhang et al., 2015), and three aspect-based sentiment classification datasets: Restaurant14 (Pontiki et al., 2016) and Laptop14 (Pontiki et al., 2016), and Twitter dataset (Dong et al., 2014).

4.2 Baselines

We compare our model with the following strong baselines for cross-domain sentiment analysis, including DANN (Ganin et al., 2016), PBLM (Ziser and Reichart, 2018), HATN (Li et al., 2018), ACAN (Qu et al., 2019), IATN (Zhang et al., 2019a) and BERT-DAAT (Du et al., 2020). BERT-DAAT is regarded as the state-of-the-art model, which uses BERT for cross-domain sentiment analysis with adversarial training. We adopt the results of these baselines reported in (Du et al., 2020). For in-domain sentiment analysis, we compare our model with SentiLR-B (Ke et al., 2019), which is one of the state-of-the-art models based on BERT.

BERT is extensively compared in our experiments. To exclude the impact of the pre-training dataset, we also compare SENTIX with BERT*, which pre-trains on the *same* dataset with standard MLM task. Moreover, to verify that SENTIX learns the sentiment knowledge in pre-training phase, we conduct our experiments with the fixed parameters of pre-trained models (marked with "$_{Fix}$"). In other words, we adopt pre-trained models as feature extractors and their parameters are not updated in fine-tuning phase.

4.3 Settings

Pre-training For pre-training phase, we use BERT as the base model and train SENTIX for 3 epochs over all the reviews data. The batch size is 256. Adam is adopted and the learning rate is set to 2e-5.

[5]https://www.yelp.com/dataset

[6]https://www.nltk.org/

	Sentiment Classification				Aspect-based Sentiment Classification		
	SST-2-Root	SST-5-Root	IMDB	Yelp	Restaurant14	Laptop14	Twitter
BERT	90.94	50.92	90.36	95.68	83.18	78.06	73.12
BERT$_{Fix}$	54.15	26.15	53.34	68.88	65.54	55.80	51.30
BERT*	90.01	51.30	90.55	95.82	84.27	78.25	73.39
BERT$^*_{Fix}$	82.98	40.36	85.03	91.68	76.25	70.10	55.85
BERT^{large}	91.47	54.80	93.53	96.72	84.20	78.53	73.12
SentiLR-B (Ke et al., 2019)	N/A	55.46	94.50	**98.03**	87.29	79.00	N/A
SentiX	92.26	54.34	92.92	97.11	86.07	79.15	**74.13**
SentiX$_{Fix}$	88.30	47.73	91.90	96.11	78.21	71.32	59.83
SentiX^{large}	**93.30**	**55.57**	**94.78**	97.83	**87.32**	**80.56**	73.99

Table 2: Results of in-domain sentiment analysis tasks.

Sentiment Analysis Similar to BERT (Devlin et al., 2019), we adopt the same settings for the downstream tasks. Specifically, we input sequence $\{[\mathrm{CLS}], w_1, ..., w_n, [\mathrm{SEP}]\}$ into pre-trained model. The last state of $[\mathrm{CLS}]$ is used as the sequence representation for classification. For aspect-based sentiment analysis, the text and aspect are concatenated as input. We search for the best random seed and learning rate (among 5e-5, 4e-5, 3e-5, and 2e-5) for BERT since it is not stable. While, SentiX is much more stable and we run it with fixed seed and learning rate (2e-5). We tune SentiX on downstream tasks with 15 epochs and keep the best model on development. Accuracy is adopted as metric for all these tasks.

5 Results and Analysis

In this section, we conduct a series of experiments to validate the performance of SentiX. First, we test our model on in-domain sentiment analysis tasks (Section 5.1) to prove that SentiX performs well for sentiment analysis. Second, to verify the effectiveness of SentiX, we conduct extensive experiments on cross-domain sentiment analysis (Section 5.2). Third, we also perform an ablation test to investigate the effectiveness of each component (Section 5.3). Forth, we explore the influence of the number of training samples (Section 5.4) and visualize the feature representations of SentiX (Section 5.5). Finally, we investigate the time complexity, space complexity, and convergence analysis of SentiX (Section 5.6).

5.1 In-Domain Sentiment Analysis

We test SentiX on two in-domain sentiment analysis tasks (sentiment classification and aspect-based sentiment classification) to verify the effectiveness of SentiX (Table 2). We observe that: **1)** SentiX performs better than the state-of-the-art model SentiLR-B in most cases, which is also based on pre-trained BERT. For Yelp, our model is not as good as (Ke et al., 2019) since it is pre-trained on the yelp review data directly. Different from (Ke et al., 2019), we pre-train our model on 30 domain datasets and the part-of-speech information is not used in our model; **2)** Compared with BERT-based baselines, SentiX obtains a significant improvement. In particular, SentiX (SentiX^{large}) performs much better than BERT (BERT^{large}) and BERT* over all seven datasets. Our pre-training objectives are effective to learn the sentiment knowledge from pre-training data; **3)** To answer how much sentiment knowledge is learned from the pre-training, we fix the parameters of SentiX, and the results show that SentiX$_{Fix}$ learns the sentiment information from large-scale dataset well, while BERT$_{Fix}$ performs only a little better than the random baseline does.

5.2 Cross-Domain Sentiment Analysis

Apart from the in-domain tasks, we conduct cross-domain sentiment analysis experiments as well. SentiX is tuned on source domain and tested on the target domain. From Table 3, we obtain the following observations:

- Compared with other works, SentiX and SentiX$_{Fix}$ achieve the best performance. SentiX$_{Fix}$ shows superior performance across all the 12 cross-domain tasks, and improves 2.56 absolute points on average over the previous the-state-of-art method (BERT-DAAT). It demonstrates that SentiX has learned the domain-invariant knowledge and transferred the sentiment knowledge from the source to

Source → Target	B→D	B→E	B→K	D→B	D→E	D→K	E→B	E→D	E→K	K→B	K→D	K→E	Avg
DANN (Ganin et al., 2016)	82.30	77.60	76.10	81.70	79.70	77.35	78.55	79.70	83.95	79.25	80.45	86.65	80.29
PBLM (Ziser and Reichart, 2018)	84.20	77.60	82.50	82.50	79.60	83.20	71.40	75.00	87.80	74.20	79.80	87.10	80.40
HATN (Li et al., 2018)	86.10	85.70	85.20	86.30	85.60	86.20	81.00	84.00	87.90	83.30	84.50	87.00	85.10
ACAN (Qu et al., 2019)	83.45	81.20	83.05	82.35	82.80	78.60	79.75	81.75	83.35	80.80	82.10	86.60	82.15
IATN (Zhang et al., 2019a)	86.80	86.50	85.90	87.00	86.90	85.80	81.80	84.10	88.70	84.70	84.10	87.60	85.90
BERT	86.75	82.80	86.20	81.55	80.60	83.00	81.85	83.85	90.80	82.10	82.05	88.35	84.13
BERT$_{Fix}$	55.40	56.55	54.05	55.10	57.25	53.75	55.50	56.00	55.55	52.30	52.75	54.15	54.86
BERT*	86.70	90.35	91.10	88.45	89.90	91.90	86.25	86.55	92.60	84.50	86.00	90.15	88.70
BERT$^*_{Fix}$	83.75	80.95	87.25	82.85	87.00	89.05	82.35	79.30	90.45	84.60	85.00	89.00	85.96
BERT-DAAT (Du et al., 2020)	89.70	89.57	90.75	90.86	89.30	87.53	88.91	90.13	93.18	87.98	88.81	91.72	90.12
SENTIX	91.15	92.50	95.70	90.85	92.15	94.95	88.10	89.86	95.45	87.00	88.05	91.85	91.47
SENTIX$_{Fix}$	**91.30**	**93.25**	**96.20**	**91.15**	**93.55**	**96.00**	**90.40**	**91.20**	**96.20**	**89.55**	**89.85**	**93.55**	**92.68**

Table 3: Experimental results of cross-domain sentiment classification. There are four domains, B: Books, D: DVD, E: Electronic, K: Kitchen & housewares. Note that we remove these four domains from pre-training data to verify the SENTIX's effectiveness of domain adaptation.

the target domain. In particular, sentiment words, emoticons and ratings from reviews are transferable signals across all domains.

- The performance of the BERT-based models is listed in the second group. BERT$_{Fix}$ only achieves 54.86% accuracy on average, which is consistent with in-domain experiments. Compared with BERT*, which is also pre-trained on the review dataset, SENTIX improves 2.8 absolute points, and we attribute it to the proposed sentiment masking and sentiment-aware pre-training objectives.

- SENTIX$_{Fix}$ performs better than SENTIX (1.21 absolute improvement on average). SENTIX$_{Fix}$ adopts the pre-trained model as feature extractor and does not update its parameters during fine-tuning, while SENTIX fine-tunes the parameters. We speculate that during fine-tuning, SENTIX learns too much domain-specific sentiment knowledge in the source domain, leading to degraded performance on the target domain. Overall, SENTIX effectively learns domain-invariant sentiment knowledge from large-scale unlabeled data and it serves as a decent sentiment feature extractor.

5.3 Ablation Study

We conduct ablation study to investigate the influence of different components from two perspectives: we remove -Sentiment, -Emoticon, and -Rating respectively to evaluate the impact of each sentiment-related pre-training task; and we remove -Token and -Sentence respectively to compare the different granularity. -Token indicates that we remove sentiment words and emoticons in the pre-training phase; and -Sentence contains only the -Rating, which excludes `RP`.

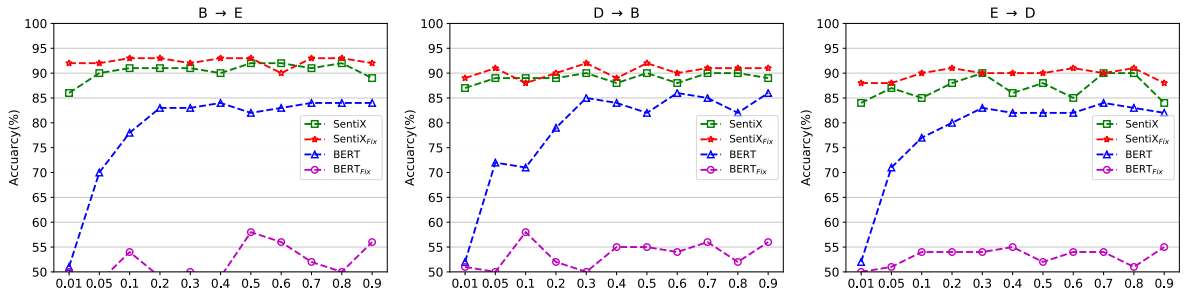
Table 4 lists the results and we observe that: **First**, each sentiment knowledge (sentiment lexicon, emoticon, and rating) improves the performance of sentiment analysis. **Second**, without rating prediction (-Rating), our model still performs better than the state-of-the-art model (BERT-DAAT). **Third**, since cross-domain sentiment analysis focuses on sentence-level sentiment, SENTIX$_{Fix}$ without sentence level objectives (-Rating) does not perform well, while SENTIX can still learn the sentence-level sentiment information from token level objectives through fine-tuning.

5.4 Influence of Sample Numbers

To study the learning curve in source domain, we test the performance of SENTIX, SENTIX$_{Fix}$, BERT and BERT$_{Fix}$ on target domain with different rates of training samples (Figure 3). **First**, we find that our model can be trained with only 1% samples (18 samples), while BERT does not work well with such limited data size. Furthermore, SENTIX with 1% samples even performs better than BERT with 90% samples. All these observations denote that SENTIX can reduce the training sample number significantly. **Second**, SENTIX$_{Fix}$ obtains better results than SENTIX, while BERT$_{Fix}$ has a poor performance. This indicates that the representation of SENTIX contains much more sentiment knowledge than the standard BERT does.

	SWM	WSP	EM	EP	RP	B → D	B → E	B → K	D → E	D → K	E → K
SOTA (BERT-DAAT)						89.70	89.57	90.75	89.30	87.53	93.18
SentiX	✓	✓	✓	✓	✓	91.15	92.50	95.70	92.15	94.95	95.45
-Sentiment			✓	✓	✓	89.97	91.65	94.15	91.25	93.20	94.75
-Emoticon	✓	✓			✓	90.37	92.01	95.05	91.33	94.05	94.67
-Token					✓	*89.15*	*90.60*	*93.35*	*90.25*	*92.25*	*93.50*
-Rating / -Sentence	✓	✓	✓	✓		90.16	91.74	95.11	91.69	94.22	94.80
SentiX$_{Fix}$	✓	✓	✓	✓	✓	91.30	93.25	96.20	93.55	96.00	96.20
-Sentiment			✓	✓	✓	90.65	91.90	95.30	92.15	95.25	95.30
-Emoticon	✓	✓			✓	90.44	92.46	94.89	92.83	94.96	95.46
-Token					✓	89.85	*91.00*	93.95	92.10	94.65	94.35
-Rating / -Sentence	✓	✓	✓	✓		*88.30*	91.05	*92.70*	*90.05*	*93.03*	*93.23*

Table 4: Ablation study on cross-domain sentiment analysis. Value marked with *underline* indicates the worst performance in the group.



(a) (b) (c)

Figure 3: The influence of sample number. We explore the influence of sample number with different rate of source domain. For the limited space, we only show the results of B → E, D → B, and E → D.

5.5 Visualization of Representation

To understand why our SentiX work, we visualize the sentence representations of BERT and SentiX for B → E task (Figure 4). In other words, the representations of source data points (books) and target data points (electronic) with positive and negative sentiment labels are provided. In particular, we convert the 768-dimensional features into two-dimension via t-SNE. From these figures, we obtain the following observations. **First**, we find that the sentences with different sentiment polarities are clearly separated in the source domain for BERT. However, some data points are mixed in the target domain. This indicates that fine-tuning the BERT overfits in the source domain. Besides, the representations of BERT$_{Fix}$ can hardly split the positive samples from the negative ones. **Second**, SentiX performs well on both of source and target domains, even it overfits in the source domain a little. The samples can be easily separated on both domains for SentiX$_{Fix}$, though the difference between positive and negative samples is not as significant as SentiX in the source domain. These demonstrate that SentiX has learned rich sentiment knowledge via pre-training tasks and avoided overfitting to a large extent.

5.6 Complexity and Convergence Analysis

In this section, we investigate the time complexity, space complexity and convergence of SentiX on B → E task (Table 5). In particular, we report the time cost for each epoch on P100 with batch size 16 and the trainable parameters in the training phase. Further, we list the accuracy on the target domain for the first five epochs to verify the convergence. We observe that SentiX$_{Fix}$ obtains better results with BERT, and SentiX$_{Fix}$ is three times as fast as BERT. Additionally, SentiX$_{Fix}$ only needs to update the classifier's parameters (2k), while the trainable parameters of BERT are much larger (133M). Moreover, our SentiX and SentiX$_{Fix}$ converge significantly faster than BERT and achieves more than 90% in the first epoch in terms of accuracy. SentiX converges with only one epoch and overfits the source domain with more epoches, while SentiX$_{Fix}$ does not overfit.

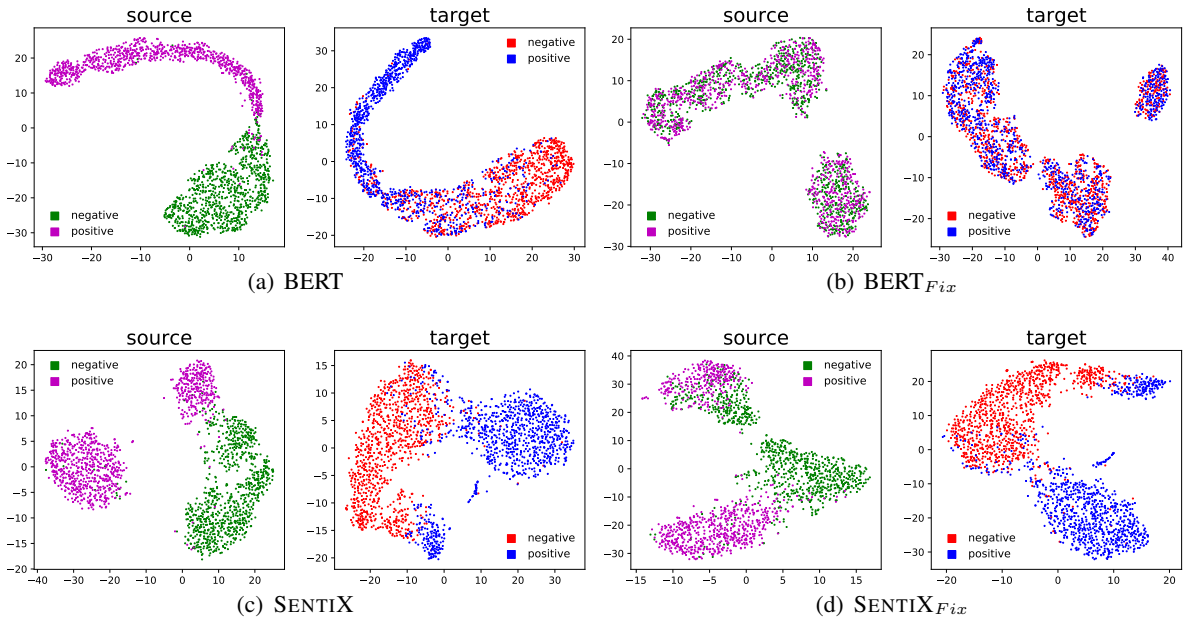


Figure 4: Visualization of sentence representation obtained from BERT and SENTIX. We use t-SNE to transfer 768-dimensional feature space into two-dimensional space for B → E task.

	Time complexity Time (Speedup)	Space complexity Trainable Parameters	Convergence (Accuracy) Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5
BERT	46s (1x)	133M	52.20	73.85	84.05	85.45	86.55
BERT$_{Fix}$	15s (3x)	2K	50.65	50.00	52.45	50.80	51.10
SENTIX	46s (1x)	133M	**92.60**	**92.30**	92.05	91.10	90.85
SENTIX$_{Fix}$	15s (3x)	2K	90.55	91.45	**92.25**	**92.75**	**93.05**

Table 5: The results of complexity and convergence. We list the costed time of each epoch and space complexity of trainable parameters in B → E task with the same batchsize.

6 Related Work

Cross-domain Sentiment Analysis Due to the heavy cost of obtaining large quantities of labeled data for each domain, many approaches have been proposed for cross-domain sentiment analysis (Blitzer et al., 2007; Yu and Jiang, 2016; Li et al., 2013; Zhang et al., 2019a; Peng et al., 2018). Most of the previous works focus on capturing the pivots that are useful for both source domain and target domain (Ziser and Reichart, 2018; Li et al., 2018). Domain adaptation adversarial training (Ganin et al., 2016) is widely-used to learn the domain-common sentiment knowledge (Li et al., 2017; Qu et al., 2019). Recently, Du et al. (2020) integrated BERT into cross-domain sentiment analysis tasks to learn the domain-shared feature representation. However, most of the existing work focuses on learning the domain-shared representation in training or fine-tuning, how to learn domain-invariant sentiment knowledge from the pre-training phase has not been explored.

Pre-trained Model Existing studies (Peters et al., 2018; Devlin et al., 2019) have proved that pre-training on large-scale unlabelled corpus obtains state-of-the-art performances in the field of natural language processing (Qiu et al., 2020). On the one hand, many studies applied pre-trained models to downstream tasks via fine-tuning (Devlin et al., 2019; Dodge et al., 2020; Sun et al., 2019b; Xu et al., 2019). Devlin et al. (2019) fine-tuned the BERT model on many downstream tasks, such as name entity recognition and sentiment analysis. Sun et al. (2019a) converted aspect-based sentiment analysis task into a sentence pair classification task to better utilize the powerful representation of BERT. On the other hand, some work proposed to add external knowledge into pre-training BERT to enhance the representations (Zhang et al., 2020). LIBERT (Lauscher et al., 2019) integrated linguistic knowledge through an additional linguistic constraint task. ERINE (Zhang et al., 2019b) and KnowBERT (Peters

et al., 2019) integrated entity representation into BERT. Alternatively, Levine et at. (2019) introduced a SenseBERT to improve lexical understanding by predicting tokens' supersenses in WordNet. Tian et al. (2020) and Ke et al. (2019) integrated external knowledge to learn sentiment information. They focused on improving the performance with fine-tuning on downstream sentiment analysis tasks by training on a relatively small or one domain dataset. Different from the existing studies, we design several pre-training objectives via rich domain-invariant sentiment knowledge in large-scale multi-domain unlabeled data for cross-domain sentiment analysis.

7 Conclusions

In this paper, we pre-train our SENTIX model to induce a general low dimensional representation based on domain-invariant sentiment knowledge for cross-domain sentiment analysis. In particular, we design several pre-training tasks to learn the sentiment knowledge from semi-supervised labels (such as sentiment words, emoticons, and ratings) based on sentiment masking. SENTIX obtains the state-of-the-art performance on 12 cross-domain sentiment analysis tasks. The visualization of the feature representation indicates that SENTIX can reduce overfitting in the source domain. The experimental results also show that SENTIX requires much less labeled data, training time and trainable parameters to obtain the equivalent performances of the standard BERT.

In the future, we are interested in exploiting more diverse pre-training datasets (e.g., twitter) and more kinds of sentiment knowledge. We also think that more self-supervised objectives could be investigated for the cross-domain sentiment analysis tasks.

Acknowledgements

The authors wish to thank the reviewers for their helpful comments and suggestions. This work was supported by National Key R&D Program of China (No. 2018AAA0100503&2018AAA0100500), and by the Science and Technology Commission of Shanghai Municipality (19511120200), Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai 200241, China. The computation is performed in ECNU Multifunctional Platform for Innovation(001). The corresponding authors are Yuanbin Wu and Liang He.

References

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.

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.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 49–54.

Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. 2020. Adversarial and domain-aware bert for cross-domain sentiment analysis. In *ACL*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. 2019. Aspect-based sentiment analysis using bert. In *NEAL Proceedings of the 22nd Nordic Conference on Computional Linguistics (NoDaLiDa), September 30-October 2, Turku, Finland*, pages 187–196.

Alexander Hogenboom, Daniella Bal, Flavius Frasincar, Malissa Bal, Franciska de Jong, and Uzay Kaymak. 2013. Exploiting emoticons in sentiment analysis. In *Proceedings of the 28th annual ACM symposium on applied computing*, pages 703–710.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Pei Ke, Haozhe Ji, Siyang Liu, Xiaoyan Zhu, and Minlie Huang. 2019. Sentilr: Linguistic knowledge enhanced language representation for sentiment analysis. *arXiv preprint arXiv:1911.02493*.

Anne Lauscher, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2019. Informing unsupervised pretraining with external linguistic knowledge. *arXiv preprint arXiv:1909.02339*.

Yoav Levine, Barak Lenz, Or Dagan, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2019. Sensebert: Driving some sense into bert. *arXiv preprint arXiv:1908.05646*.

Shoushan Li, Yunxia Xue, Zhongqing Wang, and Guodong Zhou. 2013. Active learning for cross-domain sentiment classification. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

Zheng Li, Yun Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. 2017. End-to-end adversarial memory network for cross-domain sentiment classification. In *IJCAI*, pages 2237–2243.

Zheng Li, Ying Wei, Yu Zhang, and Qiang Yang. 2018. Hierarchical attention transfer network for cross-domain sentiment classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

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

Manish Munikar, Sushil Shakya, and Aakash Shrestha. 2019. Fine-grained sentiment classification using bert. In *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, volume 1, pages 1–5. IEEE.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197.

Minlong Peng, Qi Zhang, Yu-gang Jiang, and Xuan-Jing Huang. 2018. Cross-domain sentiment classification with target domain specific information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2505–2513.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Matthew E Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *10th International Workshop on Semantic Evaluation (SemEval 2016)*.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *arXiv preprint arXiv:2003.08271*.

Xiaoye Qu, Zhikang Zou, Yu Cheng, Yang Yang, and Pan Zhou. 2019. Adversarial category alignment network for cross-domain sentiment classification. 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 2496–2508.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019a. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. 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 380–385.

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

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. Skep: Sentiment knowledge enhanced pre-training for sentiment analysis. In *ACL*.

Hu Xu, Bing Liu, Lei Shu, and S Yu Philip. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. 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 2324–2335.

Jianfei Yu and Jing Jiang. 2016. Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 236–246.

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

Kai Zhang, Hefu Zhang, Qi Liu, Hongke Zhao, Hengshu Zhu, and Enhong Chen. 2019a. Interactive attention transfer network for cross-domain sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5773–5780.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019b. Ernie: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware bert for language understanding. In *AAAI*.

Jichang Zhao, Li Dong, Junjie Wu, and Ke Xu. 2012. Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1528–1531.

Yftah Ziser and Roi Reichart. 2018. Pivot based language modeling for improved neural domain adaptation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1241–1251.