Diversified Multiple Instance Learning for Document-Level Multi-Aspect Sentiment Classification

Yunjie Ji[1,2], Hao Liu[1], Bolei He[1], Xinyan Xiao[1]*, Hua Wu[1], Yanhua Yu[2]
[1]Baidu Inc., Beijing, China [2]Beijing University of Posts and Telecommunications
{jiyunjie,liuhao24,hebolei,xiaoxinyan,wu_hua}@baidu.com
yuyanhua@bupt.edu.cn

Abstract

Neural Document-level Multi-aspect Sentiment Classification (DMSC) usually requires a lot of manual aspect-level sentiment annotations, which is time-consuming and laborious. As document-level sentiment labeled data are widely available from online service, it is valuable to perform DMSC with such free document-level annotations. To this end, we propose a novel Diversified Multiple Instance Learning Network (D-MILN), which is able to achieve aspect-level sentiment classification with only document-level weak supervision. Specifically, we connect aspect-level and document-level sentiment by formulating this problem as multiple instance learning, providing a way to learn aspect-level classifier from the back propagation of document-level supervision. Two diversified regularizations are further introduced in order to avoid the overfitting on document-level signals during training. Diversified textual regularization encourages the classifier to select aspect-relevant snippets, and diversified sentimental regularization prevents the aspect-level sentiments from being overly consistent with document-level sentiment. Experimental results on TripAdvisor and BeerAdvocate datasets show that D-MILN remarkably outperforms recent weakly-supervised baselines, and is also comparable to the supervised method.

1 Introduction

Document-level multi-aspect sentiment classification (DMSC) is a fine-grained sentiment analysis task, aiming to predict the sentiments of aspects in a document consisting of several sentences. In previous studies, neural models have shown to be effective for improving DMSC with the help of large amounts of aspect-level annotations (Chen et al., 2017; Xue and Li, 2018; Chen and Qian, 2019; Wang et al., 2020). Despite the advantages, the acquisition of aspect-level sentiment annotations remains a laborious and expensive endeavor. Fortunately, the overall document-level sentiment annotations are relatively easy to obtain thanks to the widespread online reviews with overall star ratings. Therefore, it is practically meaningful to perform DMSC by weak supervision from document-level sentiment signals.

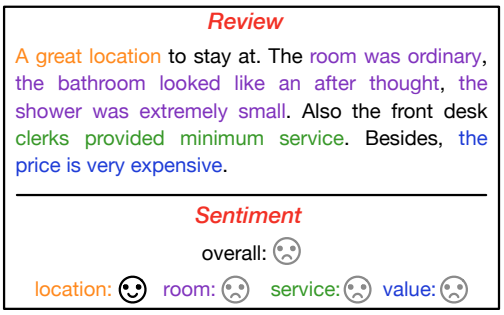


Figure 1: A review example with sentiment labels.

However, this problem is far from solved. To the best of our knowledge, there is no neural model that is able to achieve DMSC with only document-level signals. There are mainly two challenges need to be settled. First, the granularity between aspect-level sentiment and document-level sentiment is quite different. It is unclear how to properly model the relation between them, in order to transfer knowledge from document-level to aspect-level. Second, the relevant text of aspect-level is unobserved. Without any constraint, a vanilla weakly supervised model would be easy to overfit to document-level signals in terms of both sentiment and attended text, despite each aspect often has its unique relevant text and different sentiment (as shown in Figure 1). However in this case, no matter the given aspect is *location*, *room*, *service*, or *value*, a vanilla model would pay more attention to the words "great", "ordinary", "small", "minimum" and "expensive",

* corresponding author.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 7012–7023,
November 16–20, 2020.

and transfer the negative sentiment from document-level to all aspects. As a result, the sentiment towards *location* is wrongly learned as negative, which should be positive instead.

Accordingly, we propose a *diversified multiple instance learning network* (D-MILN) to achieve DMSC with only document-level sentiment supervision. We novelly formularize this problem as *multiple instance learning* (MIL; Keeler and Rumelhart 1991) to model document-level sentiment as a combination of aspect-level sentiments. The aspects are regarded as instances and their sentiment distributions are predicted by an attention-based classifier, while the document is regarded as a bag and its sentiment distribution is computed as a combination of the aspect-level sentiment distributions. Thus, we provide a framework for learning aspect-level classifier by optimizing the document-level predictions. Meanwhile, in order to avoid the overfiting to document-level signals, we further propose two kinds of *diversified regularization*. Diversified textual regularization is applied to guide the aspect-level sentiment classifier to select aspect-relevant snippets. Diversified sentimental regularization is leveraged to control the variance among aspect-level sentiments. Overall, our contributions are summarized as follows:

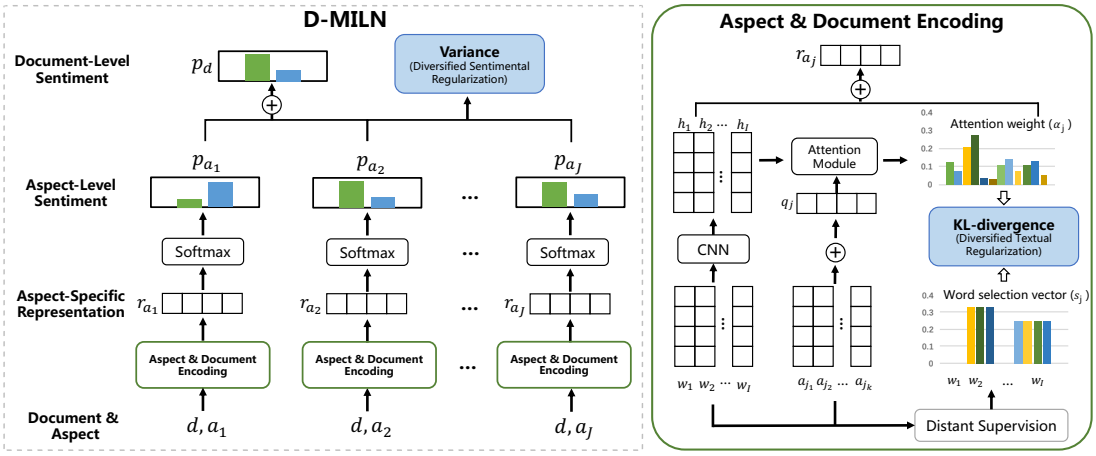
- We propose a novel diversified multiple instance learning neural network, which properly models the relation between aspect-level and document-level sentiment, and thus achieves DMSC with merely document-level supervision.
- Two kinds of diversified regularization are introduced to alleviate the key challenge of overfitting document-level signals and to improve the aspect-level sentiment classification performance.
- Comprehensive experiments are conducted on the BeerAdvocate and TripAdvisor benchmark datasets. The results verify the necessity and advantages of both our framework and diversified regularizations. Meanwhile, our D-MILN outperforms previous weakly supervised methods significantly and is also comparable to the supervised method with thousands of labeled instances per aspect.

2 Related Work

Document-level multi-aspect sentiment classification In previous studies, DMSC is usually done by supervised learning methods (Lei et al., 2016; Yin et al., 2017; Li et al., 2018; Wang et al., 2019), where aspect-level annotations should be provided. However, human annotation of aspect-level sentiment is laborious and expensive, therefore, some researches focus on weakly supervised DMSC. This approach can be further categorized into knowledge-supervised and document-level supervised methods. As for knowledge-supervised methods, Zeng et al. (2019) propose to use aspect-opinion word pairs as knowledge for supervision. The aspect-level sentiment classification is achieved by accomplishing another relevant objective: to predict an opinion word when given an aspect. However, their model heavily depends on the performance of dependency parsing and manually designed rules. As for document-level supervised methods, Wang et al. (2010, 2011) propose to use the document-level sentiment as supervision which is similar to ours. Specifically, they propose a probabilistic graphical model for the task, which assumes the overall rating is generated based on a weighted sum of the latent aspect ratings. However, this non-neural network model adopts bag-of-words representations which are insufficient at capturing the order of words and complex semantics. Furthermore, their model fails to consider the problem of overfitting to document-level signals.

Multiple Instance Learning Multiple instance learning is a form of weakly supervised learning where instances are arranged in bags and a label is provided for the entire bag (Keeler and Rumelhart, 1991). Most MIL methods (Zhou et al., 2009; Wei et al., 2014; Pappas and Popescu-Belis, 2017; Haußmann et al., 2017; Tu et al., 2019; Ilse et al., 2018; Wang et al., 2018; Wang and Wan, 2018) focus on the bag-level performance and there are also a few methods focusing on the instance-level performance. Apart from the loss defined on the bag level, Kotzias et al. (2015) also introduces a regularization based on the instance similarities into the objective function. Peng and Zhang (2019) assigns the bag-level label to instances under the i.i.d assumption and directly define the loss function on the instance-level label prediction.

Some works propose to apply MIL to sentence-level sentiment classification task. Kotzias et al. (2015); Angelidis and Lapata (2018a); Wang and



(a) The main architecture of D-MILN

(b) The aspect & document encoding module

Figure 2: Architecture of our Diversified Multiple Instance Learning Network (D-MILN).

Wan (2018) and Angelidis and Lapata (2018b) propose to train the sentence-level sentiment classifier with document-level annotations. For these works, the content for each instance (i.e. words in the sentence) is already given. However, for DMSC task, the relevant text snippets for a given aspect, which are crucial for determining the sentiment, are not provided in advance. This makes the DMSC task much different and challenging to apply MIL. Besides, these works never consider the overfitting to bag-level supervision. To the best of our knowledge, this is the first work to apply MIL to DMSC task.

3 Methodology

We first briefly introduce the problem we work on. Given a review, our task is to predict the sentiments of aspects in the review. Formally, we denote the review document as d which contains I words $\{w_1, w_2, \cdots, w_I\}$, the sentiment label for the document as l_d, and the set of J aspects mentioned in the document as $\{a_1, a_2, \cdots, a_J\}$. Same as Yin et al. (2017), each aspect a_j is represented by K aspect-related keywords, $\{a_{j_1}, a_{j_2}, \cdots, a_{j_K}\}$, in order to cover most of the semantic meanings of the aspect[1].

Figure 2 shows the architecture of D-MILN, where Figure 2(a) is the entire workflow and Figure 2(b) is the detailed network of aspect and document encoding. First, the aspect-level attention-based classifier predicts sentiment distributions for every mentioned aspect which are denoted as $p_{a_1}, p_{a_2}, \cdots, p_{a_J}$. Then, the document-level sentiment distribution p_d is computed as a weighted sum of aspect-level sentiment distributions. The diversified sentimental regularization as shown in Figure 2(a) is applied on the aspect-level sentiment distributions to alleviate the overfitting to document-level sentiment. The diversified textual regularization as shown in Figure 2(b) is applied on the attention weights to encourage the aspect-level classifier to select aspect-relevant snippets.

3.1 Aspect-level Sentiment Distribution

In this section, we introduce our aspect-level attention-based sentiment classifier.

Aspect encoding We first apply a one-layer MLP on the top of word embedding of each aspect-related keyword a_{j_k}:

$$\mathbf{q}_{j_k} = tanh(\mathbf{W}_q \mathbf{e}_{j_k} + \mathbf{b}_q) \tag{1}$$

where $\mathbf{e}_{j_k}$ is the word embedding of a_{j_k}, $\mathbf{W}_q$ and $\mathbf{b}_q$ are parameters of the one-layer MLP. Then the final representation of aspect a_j is calculated as $\mathbf{q}_j = \sum_k c_k \mathbf{q}_{j_k}$, where c_k encodes the importance of each keyword for the given aspect:

$$c_k = \frac{\exp(\mathbf{w}_c \cdot \mathbf{q}_{j_k})}{\sum_{k'} \exp(\mathbf{w}_c \cdot \mathbf{q}_{j_{k'}})} \tag{2}$$

and $\mathbf{w}_c$ is the parameter to learn.

Document encoding We first convert the words in the given document into a sequence of embedding vectors $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_I]$. Usually, the sentiments are expressed through phrases in the document (Fei et al., 2004). For example, "a lovely room" expresses a positive sentiment towards the aspect *room*. Since one-dimension convolutional

[1] See Appendix A.1 for the keywords.

layers can serve as linguistic feature detectors to extract specific patterns of n-grams (Kalchbrenner et al., 2014), we apply several one-dimension convolutional layers on top of the word embeddings and obtain the final contextual features for the input words: $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_I]$.

Aspect-specific representations We obtain the aspect-specific representation by a weighted sum of contextual features:

$$\mathbf{r}_{a_j} = \sum_{i=1}^{I} \alpha_j^i \mathbf{h}_i \tag{3}$$

where α_j^i encodes the importance of word w_i to determine the sentiment towards aspect a_j. α_j^i is calculated through attention mechanism:

$$\alpha_j^i = \frac{\exp({\mathbf{q}_j}^T \mathbf{W}_a \mathbf{h}_i)}{\sum_{i'} \exp({\mathbf{q}_j}^T \mathbf{W}_a \mathbf{h}_{i'})} \tag{4}$$

where $\mathbf{W}_a$ is a bilinear term to capture the relevance between $\mathbf{q}_j$ and $\mathbf{h}_i$.

Prediction The aspect-specific representation is then used to predict the aspect-level sentiment distribution p_{a_j} by:

$$p_{a_j} = softmax(\mathbf{W}_p \mathbf{r}_{a_j} + \mathbf{b}_p) \tag{5}$$

where $\mathbf{W}_p$ and $\mathbf{b}_p$ are parameters of the softmax layer.

3.2 Document-level Sentiment Distribution

Since only document-level supervision is provided, we could not directly use the aspect-level sentiment distribution p_{a_j} for optimization. In order to connect aspect-level sentiment with document-level sentiment, we compute document-level sentiment distribution as a weighted sum of aspect-level distributions. Thus, by optimizing the document-level predictions, the parameters of the aspect-level sentiment classifier are learned through back propagation. Specifically, the document-level distribution is as following:

$$p_d = \sum_{j=1}^{J} \beta_j p_{a_j} \tag{6}$$

where β_j encodes the importance of aspect a_j for determining the sentiment of the overall document. To obtain β_j, we first average the aspect representations:

$$\mathbf{r}_d = \frac{1}{J} \sum_{j=1}^{J} \mathbf{r}_{a_j} \tag{7}$$

then we use attention mechanism to derive β_j:

$$\beta_j = \frac{\exp(\mathbf{v}_r^T tanh(\mathbf{W}_r[\mathbf{r}_{a_j}; \mathbf{r}_d] + \mathbf{b}_r))}{\sum_{j'} \exp(\mathbf{v}_r^T tanh(\mathbf{W}_r[\mathbf{r}_{a_{j'}}; \mathbf{r}_d] + \mathbf{b}_r))} \tag{8}$$

where $[\mathbf{r}_{a_j}; \mathbf{r}_d]$ is the concatenation of $\mathbf{r}_{a_j}$ and $\mathbf{r}_d$, $\mathbf{W}_r$, $\mathbf{b}_r$ and $\mathbf{v}_r$ are parameters of the attention mechanism. .

After obtaining document-level sentiment distributions, we train the model with respect to document-level sentiment labels and introspectively, the aspect-level sentiment classifier is learned through back propagation.

3.3 Diversified Regularizations

The aspect-level sentiment classifier simply learned in such a way suffers from the overfitting to document-level supervision signals. Firstly, given different aspects, the aspect-level sentiment classifier tends to focus on the same snippets, which actually express the document-level sentiment. Secondly, the predicted aspect-level sentiments tend to be overly consistent with the document-level sentiment.

Diversified Textual Regularization To alleviate the first problem, diversified textual regularization is proposed to encourage the sentiment classifier to select aspect-relevant snippets with distant supervision. The main idea is that the aspect-level classifier should pay more attention to the words which co-occur with the given aspect in a same sentence. Specifically, given an aspect a_j, a distantly-labeled word selection vector $\mathbf{s}_j$ is leveraged to guide the attention weight vector $\boldsymbol{\alpha}_j$ in Equation 4. To obtain $\mathbf{s}_j$, we first initialize the weights of all words in the document to be 0. Secondly, we find the sentences which contain any keywords of the given aspect[2]. Then we set the weights of words in these sentences to be 1. Finally, we normalize the weight vector. The diversified textual regularization is defined as the KL-divergence betweens $\boldsymbol{\alpha}_j$ and $\mathbf{s}_j$:

$$L_{d-text} = KL(\mathbf{s}_j || \boldsymbol{\alpha}_j) = \sum_i \mathbf{s}_j^i \log \frac{\mathbf{s}_j^i}{\boldsymbol{\alpha}_j^i} \tag{9}$$

Furthermore, there exist sentences which describe multiple aspects. As in most of these sentences, the parts related to different aspects are

[2]We experiment with different levels of snippets (sentence-level and clause-level). Experimental results show that sentence-level snippets achieve more promising results. We guess that is because the clause-level snippets may be incomplete or biased for expressing sentiments.

Dataset	*#docs*	*#words/doc*	*#aspects*	*#mentioned-aspects*	*#one-aspect-docs*	*#aspect-labels*
TripAdvisor	17,792	251.7	7	4.7	20/0.07%	11,915
BeerAdvocate	15,338	144.5	4	3.3	877/3.18%	12,686

Table 1: Statistics of two datasets. *#docs* denotes the number of training documents. *#words/doc* is the average number of words in each document. *#aspects* denotes the number of predefined aspects, while *#mentioned-aspects* is the average number of aspects mentioned in each document. *#one-aspect-docs* denotes the number/percentage of documents in which only one aspect is mentioned. *#aspect-labels* denotes the average number of labeled instances for each aspect.

non-overlapping, we also apply orthogonal regularization (Lin et al., 2017; Hu et al., 2018) to guide the attention weights in a fine granularity:

$$L_{ortho} = \sum_{j} \sum_{j' \neq j} \boldsymbol{\alpha}_j \cdot \boldsymbol{\alpha}_{j'} \tag{10}$$

Minimizing the dot product between two attention weight vectors will force orthogonality between them, so that different aspects attend on different parts of the sentence with less overlap.

Diversified Sentimental Regularization Given a document, some of its aspects often have different sentiments from the document-level sentiment. But simply fitting the document-level supervision leads the sentiments of all aspects to be same with the document-level sentiment. To tackle this problem, we propose diversified sentimental regularization to control the variance among aspect-level sentiment distributions. The variance is computed as follows:

$$\begin{aligned} L_{d-senti} &= \frac{1}{J} \sum_{j=1}^{J} (p_{a_j}(l_d) - p_u(l_d))^2 \\ p_u(l_d) &= \frac{1}{J} \sum_{j=1}^{J} p_{a_j}(l_d) \end{aligned} \tag{11}$$

where $p_{a_j}(l_d)$ is the probability of class l_d for aspect a_j. By maximizing $L_{d-senti}$, the model allows the aspect-level sentiment distributions to be different, so that for some aspects, their sentiments could be different from the document-level sentiment l_d. Furthermore, instead of using cross-entropy loss, we propose to leverage hinge loss to control the fitting degree of the document-level sentiment distribution p_d to the ground truth label l_d. The hinge loss is defined as follows:

$$L_{doc} = max(t - p_d(l_d), 0) \tag{12}$$

where $p_d(l_d)$ is the probability of the ground-truth label l_d, $t \in (0.5, 1.0]$ is the probabilistic margin, which gives the tolerance to diverse aspect-level sentiment distributions.

3.4 Final Objective Function

The final objective function of D-MILN is a combination of document-level loss and diversified regularizations. To minimize clutter, we describe the objective function for a single document:

$$L = L_{doc} + \alpha^m L_{d-text} + \beta L_{ortho} + \gamma L_{d-senti} \tag{13}$$

where α, β, γ are the hyper-parameters, m is the number of training steps. In diversified textual regularization, the distant supervision is relatively "hard" on the attention weights, which may hurt the generalization of D-MILN, so we further introduce a decay factor $\alpha \in (0, 1)$. With the increase of training steps (m), the weight of textual diversified regularization will decrease to zero such that the model will be allowed to achieve better generalization. γ controls the sentimental diversity among aspects. For $\gamma < 0$, the sentimental diversity is encouraged. For $\gamma > 0$, the sentimental diversity is discouraged.

4 Experiments

4.1 Datasets

We evaluate our model on TripAdvisor (Wang et al., 2010) and BeerAdvocate (McAuley et al., 2012) benchmark datasets, which contain seven predefined aspects (*value*, *room*, *location*, *cleanliness*, *check in/front desk*, *service*, and *business*) and four predefined aspects (*feel*, *look*, *smell*, and *taste*) respectively. We run the same preprocessing steps as Zeng et al. (2019). The original ratings of TripAdvisor and BeerAdvocate datasets are converted to binary scales, namely, positive or negative. The exploration on fine-grained sentiment classification remains for future work. The number of reviews with negative overall sentiment and that with positive overall sentiment are balanced. Table 1 shows the statistics of the two datasets. Both datasets are split into train/development/test sets with proportions 8:1:1. The development set is used to tune the hyper-parameters for all methods. We use accuracy

as the evaluation metric. Note that both aspect-level and document-level sentiment annotations are provided in the datasets, but our D-MILN only uses document-level annotations for training.

4.2 Implementation Details

We adopt the pre-trained uncased GloVe 300-dimensional word embeddings (Pennington et al., 2014), which are set to be trainable during the training process[3]. In document encoding, we apply three one-dimension convolutional layers with kernel widths of 3, 5, and 7 respectively[4]. The number of filters is 200 for each convolutional layer. Batch normalization is applied on the output of the convolutional layers. The dimension of all hidden layers is 200. Dropout is applied on the embedding layer and the final representations of aspects and document words with dropout rate being 0.4. The values of α, β, γ in Equation 13 are 0.999, 0.1 and -0.1 respectively. The probabilistic margin t is 0.7. The batch size is set to be 64. Parameter optimization is performed using Adam (Kingma and Ba, 2014) with learning rate being 0.001. We run experiments on one Tesla V100 16GB GPU and each epoch takes several minutes. Our model has 438K parameters, not including word embeddings.

4.3 Compared Methods

Here, we compare our method with a variety of baselines, which can be divided into three categories. (1) Weakly supervised baselines. We use these baselines to show the advancement of D-MILN in terms of weak supervision. (2) MIL baselines. We novelly formulate weakly supervised DMSC as MIL for the first time. By comparing with several simple MIL methods, we also hope to see the necessity of D-MILN. (3) Supervised baseline. Finally, we compare D-MILN with supervised baselines to analyse the performance gap with supervised methods.

4.3.1 Weakly Supervised Baselines

Assign-O, which directly uses the overall sentiment of a review in the test set as the prediction for its aspects.

LRR (Wang et al., 2010), which is a probabilistic graphical model (non-neural model) that regards the aspect-level sentiments as latent variables and assumes the document-level sentiment is generated based on a weighted sum of the latent aspect sentiments. LRR only requires document-level annotations.

VWS-DMSC (Zeng et al., 2019), which is previous state-of-the-art weakly supervised approach for DMSC. VWS-DMS uses aspect-opinion word pairs as supervision. The sentiment of an aspect is treated as a latent variable and is used to predict the opinion word of the given aspect. VWS-DMSC also uses document-level sentiment labels to train a document encoder.

4.3.2 MIL Baselines

Vanilla-MILN, which is derived by removing key components from D-MILN. Specifically, in Vanilla-MILN, the loss function is cross-entropy loss and the diversified regularizations are not applied.

Identity-MILN, which sets the aspect-level sentiment of training data to be identical with document-level labels, and directly trains the aspect-level attention-based sentiment classifier introduced in Section 3.1.

Explicit-MILN, of which the relevant snippets for each aspect are firstly extracted by an iterative method adopted in Wang et al. (2010), then a CNN-based text classifier is applied on the extracted snippets to predict the aspect-level sentiment under the MIL framework.

4.3.3 Supervised Baselines

AB-DMSC, which is the attention-based aspect-level sentiment classifier introduced in Section 3.1. We directly train this classifier with entire aspect-level sentiment annotations. AB-DMSC serves as an upper bound to our model.

AB-DMSC-{500, 1000, 2000, 5000}, which is the AB-DMSC model trained with {500, 1000, 2000, 5000} labeled instances **per aspect**. Since the sampled labeled data may vary for different trials, we perform five trials of random sampling and report both mean and standard deviation of the results.

N-DMSC (Yin et al., 2017), which is the state-of-the-art supervised neural model. N-DMSC is also trained with entire aspect-level sentiment annotations.

[3] We use pre-trained word embeddings rather than BERT, because we find that BERT is easy to be overfitting in this problem and produces worse results. See Appendix A.2 for more details.

[4] We test a lot of combinations and find that the performance is better with bigger kernel widths. For most documents, especially of BeerAdvocate, there is almost no single word directly expressing the sentiment towards an aspect, so the model should focus on the pattern of a wide range of words to determine the sentiment.

Model	TripAdvisor		BeerAdvocate	
	Mean	Std	Mean	Std
Assign-O†	0.7043	-	0.6570	-
LRR†	0.6947	0.0024	0.5941	0.0113
VWS-DMSC†	**0.7561**	0.0012	**0.7538**	0.0066
Vanilla-MILN	0.7163	-	0.7250	-
Identity-MILN	0.7420	-	0.7124	-
Explicit-MILN	0.7618	-	0.7591	-
D-MILN (Our)	**0.7952**	-	**0.7986**	-
AB-DMSC-500	0.7566	0.0030	0.7518	0.0031
AB-DMSC-1000	0.7674	0.0042	0.7715	0.0015
AB-DMSC-2000	0.7941	0.0028	0.8009	0.0021
AB-DMSC-5000	0.8211	0.0016	0.8389	0.0031
AB-DMSC	**0.8374**	-	0.8598	-
N-DMSC†	0.8334	-	**0.8635**	-

Table 2: Averaged accuracies on the two datasets. The standard deviation is also reported for methods involving randomness during training. The maximum accuracy in each block is highlighted in bold. † : The results from Zeng et al. (2019).

4.4 Results and Analysis

Table 2 shows the main results. It contains three blocks, corresponding to the three categories of systems.We compare D-MILN with them as follows.

(1) Weakly Supervised Baselines. Our model achieves the best performance comparing with previous weakly supervised baselines. From Assign-O, we can see that directly transferring the document-level sentiment to aspects gives a poor result, showing the difficulty and necessity of finding a way to properly model the relation between document-level sentiment and aspect-level sentiment. Our model outperforms the traditional probabilistic graphical model LRR with a substantial margin, which demonstrates the necessity of utilizing neural networks to capture deep semantic features. Our model also outperforms previous SOTA VWS-DMSC significantly. VWS-DMSC relies on the extracted aspect-opinion word pairs, but we find that there are no typical opinion words for some aspects in the corpus (e.g. *look* in BeerAdvocate). Besides, in VWS-DMSC, the document-level supervision is only used to train a document encoder, which ignores the relationship between aspects and documents. As our D-MILN only relies on document-level signals, this further confirms that D-MILN properly models the relation between aspect-level and document-level sentiment.

(2) MIL Baselines. D-MILN significantly outperforms all MIL baselines with a substantial margin. Meanwhile, we find simple MIL baselines often fail to improve performance against previous work (LRR and VWS-DMSC), showing the difficulty of achieving weakly-supervised DMSC by MIL. Furthermore, from Vanilla-MILN, we can conclude that locating aspect-relevant snippets and overcoming the overfitting to document-level supervision are two challenges to improve the performance of MIL on DMSC. Compared with Identity-MILN, it suggests that our method could reduce the noises brought from the document-level supervision signals. Compared with Explicit-MILN, it suggests that our method could effectively select aspect relevant snippets.

(3) Supervised Baselines. we first find that AB-DMSC is comparable with N-DMSC, which demonstrates that our aspect-level sentiment classifier could serve as a strong supervised baseline model. Our D-MILN is comparable with AB-DMSC-2000. To analyse the performance gap between D-MILN and AB-DMSC, we conduct a case study, which is contained in Appendix A.3, to qualitatively evaluate the aspect-level attention-based sentiment classifiers.

Model	TripAdvisor	BeerAdvocate
D-MILN	0.7952	0.7986
– keywords	0.7866	0.7878
– orthogonal	0.7842	0.7955
– hinge loss	0.7795	0.7881
– d-senti	0.7631	0.7702
– d-text	0.7172	0.6742

Table 3: Accuracies on the two datasets in the ablation study.

4.5 Ablation Study

To demonstrate the effectiveness of each component of D-MILN, we conduct an ablation study and list the results in Table 3. "– keywords" means simply using the aspect term rather than its keywords to interact with the document. "– hinge loss" means replacing the hinge loss in Equation 12 by cross-entropy loss. "– d-senti" means removing diversified sentimental regularization. "– d-text" means removing diversified textual regularization. We can see that extending a single aspect term with a list of aspect relevant keywords can improve the classification performance on both datasets. The orthogonal regularization is much more useful in the TripAdvisor dataset, which indicates there are more sentences containing multiple aspects. By employing the diversified sentimental regularization, the overfitting problem of document-level signals can be alleviated and thus improves the classification performance. When removing the diversified

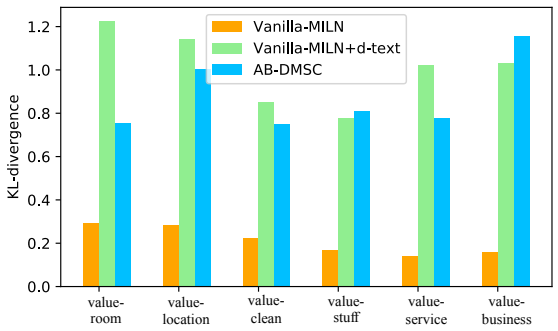


Figure 3: The KL-divergences between attention weight distributions of different aspect pairs.

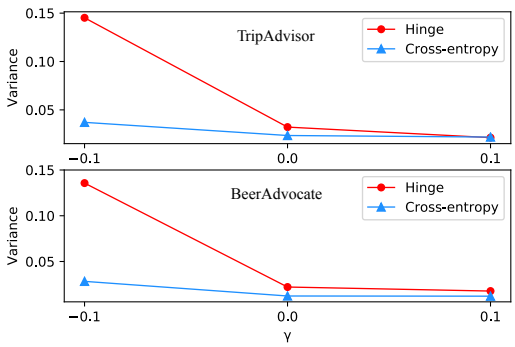


Figure 4: The variances among aspect-level sentiment distributions with different loss functions.

textual regularization, the results are much worse than removing other components, demonstrating locating the aspect-relevant snippets is crucial for correctly predicting the aspect-level sentiments.

4.6 Effectiveness of Diversified Textual Regularization

To further demonstrate the effectiveness of diversified textual regularization, we display the KL-divergence between attention weight distributions of different aspect pairs in Figure 3. The attention weight distribution, which is calculated by Equation 4, indicates the importances of document words to the given aspect. Large KL-divergences indicate that the aspect-level classifier selects distinct snippets for different aspects. For Vanilla-MILN, the KL-divergences are relatively small, which indicates that the model focuses on similar snippets for different aspects. For Vanilla-MILN+d-text, on which the diversified textual regularization is applied, the KL-divergences become larger and are similar with that of AB-DMSC, which is trained with aspect-level annotations and produces the most proper attention weights among the three models. Such results indicate that diversified textual regularization encourages the aspect-level sentiment classifier to select aspect-relevant snippets.

4.7 Hinge Loss for Diversified Sentimental Regularization

We further demonstrate that hinge loss is more compatible than cross-entropy loss with diversified sentimental regularization. In Figure 4, we display the variances, which is calculated by Equation 11, among aspect-level sentiment distributions when different loss functions are adopted. The horizontal axis γ denotes the weight of the diversified sentimental regularization. When γ turns to 0.0, which means the diversified sentimental regularization is not applied, we find that the variance is relatively small for both hinge loss and cross entropy loss, which indicates that the predicted aspect-level sentiments are over consistent with document-level ones. When γ turns to -0.1, which means the diversity of sentiments is encouraged, the variance under hinge loss grows significantly than cross-entropy loss, which verifies that by applying hinge loss, the diversity among aspect-level sentiments could be controlled more effectively.

5 Conclusion

In this paper, we propose a *diversified multiple instance learning network* to achieve DMSC with only document-level supervision. We formulate this problem as multiple instance learning, so as to model the relation between aspect-level sentiment and document-level sentiment. In order to guarantee the proper transfer from document-level supervision to aspect-level prediction, we further propose diversified textual regularization and diversified sentimental regularization. Through experiments on two benchmark datasets, we verify that our D-MILN can properly capture the interaction between aspect-level and document-level, and achieve new SOTA on weakly supervised DMSC. Detailed comparisons also show the necessity and effectiveness of our diversified regularizations. In the future, we plan to further improve D-MILN with aspect-level annotations and find appropriate way to combine D-MILN with pre-training methods (Tian et al., 2020).

Acknowledgments

This work was supported by the National Key Research and Development Project of China (No. 2018AAA0101900) and National Natural Science Foundation of China (No. U1936104).

References

Stefanos Angelidis and Mirella Lapata. 2018a. Multiple instance learning networks for fine-grained sentiment analysis. *Transactions of the Association for Computational Linguistics*, 6:17–31.

Stefanos Angelidis and Mirella Lapata. 2018b. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. *arXiv preprint arXiv:1808.08858*.

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 452–461.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhongchao Fei, Jian Liu, and Gengfeng Wu. 2004. Sentiment classification using phrase patterns. In *The Fourth International Conference onComputer and Information Technology, 2004. CIT'04.*, pages 1147–1152. IEEE.

Manuel Haußmann, Fred A Hamprecht, and Melih Kandemir. 2017. Variational bayesian multiple instance learning with gaussian processes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6570–6579.

Mengting Hu, Shiwan Zhao, Li Zhang, Keke Cai, Zhong Su, Renhong Cheng, and Xiaowei Shen. 2018. Can: Constrained attention networks for multi-aspect sentiment analysis. *arXiv preprint arXiv:1812.10735*.

Maximilian Ilse, Jakub M Tomczak, and Max Welling. 2018. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

James D. Keeler and David E. Rumelhart. 1991. A self-organizing integrated segmentation and recognition neural net. In *Advances in Neural Information Processing Systems 4, [NIPS Conference, Denver, Colorado, USA, December 2-5, 1991]*.

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

Dimitrios Kotzias, Misha Denil, Nando De Freitas, and Padhraic Smyth. 2015. From group to individual labels using deep features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 597–606. ACM.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*.

Junjie Li, Haitong Yang, and Chengqing Zong. 2018. Document-level multi-aspect sentiment classification by jointly modeling users, aspects, and overall ratings. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 925–936.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining*, pages 1020–1025. IEEE.

Nikolaos Pappas and Andrei Popescu-Belis. 2017. Explicit document modeling through weighted multiple-instance learning. *Journal of Artificial Intelligence Research*, 58:591–626.

Minlong Peng and Qi Zhang. 2019. Address instance-level label prediction in multiple instance learning. *arXiv preprint arXiv:1905.12226*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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 e-prints*.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4067–4076, Online. Association for Computational Linguistics.

Ming Tu, Jing Huang, Xiaodong He, and Bowen Zhou. 2019. Multiple instance learning with graph neural networks. *arXiv preprint arXiv:1906.04881*.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 783–792. ACm.

Hongning Wang, Yue Lu, and ChengXiang Zhai. 2011. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 618–626. ACM.

Jingjing Wang, Changlong Sun, Shoushan Li, Jiancheng Wang, Luo Si, Min Zhang, Xiaozhong Liu, and Guodong Zhou. 2019. Human-like decision making: Document-level aspect sentiment classification via hierarchical reinforcement learning. 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 5585–5594, Hong Kong, China. Association for Computational Linguistics.

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020. Relational graph attention network for aspect-based sentiment analysis. *arXiv preprint arXiv:2004.12362*.

Ke Wang and Xiaojun Wan. 2018. Sentiment analysis of peer review texts for scholarly papers. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 175–184. ACM.

Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. 2018. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24.

Xiu-Shen Wei, Jianxin Wu, and Zhi-Hua Zhou. 2014. Scalable multi-instance learning. In *2014 IEEE International Conference on Data Mining*, pages 1037–1042. IEEE.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2514–2523, Melbourne, Australia. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Yichun Yin, Yangqiu Song, and Ming Zhang. 2017. Document-level multi-aspect sentiment classification as machine comprehension. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2044–2054.

Ziqian Zeng, Wenxuan Zhou, Xin Liu, and Yangqiu Song. 2019. A variational approach to weakly supervised document-level multi-aspect sentiment classification. *arXiv preprint arXiv:1904.05055*.

Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. 2009. Multi-instance learning by treating instances as non-iid samples. In *Proceedings of the 26th annual international conference on machine learning*, pages 1249–1256. ACM.

A Appendices

A.1 Aspect-related keywords

The aspect-related keywords are listed in Table 4. For TripAdvisor, 10 keywords are provided for each aspect. For BeerAdvocate, 5 keywords are provided for each aspect.

A.2 Pre-trained Model in MIL

We are also curious about the application of pre-trained models (e.g. BERT (Devlin et al., 2018)) in MIL, since they have achieved a great success in many NLP tasks (Raffel et al., 2019; Yang et al., 2019; Lan et al., 2019). However, we find that BERT finetuned with document-level supervision is more likely to overfit the document-level sentiment supervision and thus the performance on aspect-level sentiment prediction degrades. We propose four ways to apply BERT in MIL:

BERT-asp: We first fine tune BERT with the aspect-level annotations to demonstrate the superiority of it in supervised aspect-level sentiment classification task. Specifically, the list of keywords of the given aspect is regarded as"text A", the given document is regarded as "text B", and the sentiment distribution of the aspect is calculated on the final representation of token [CLS].

BERT-doc: To adapt BERT to MIL, we combine aspects' sentiment distributions which are obtained in the same way as BERT-asp to form the document-level sentiment distribution. Then we train the model only with respect to the document-level annotations.

BERT-enc-fix: We replace the CNN encoder of D-MILN with BERT (i.e. treat BERT as a feature extractor) and set the parameters of BERT to be fixed during training.

BERT-enc-train: We replace the CNN encoder of D-MILN with BERT and set the parameters of BERT to be trainable during training.

From table 5, We can see that BERT-asp outperforms N-DMSC significantly, producing new state-of-the-art results. By comparing BERT-asp and

TripAdvisor	Keywords
value	value, price, quality, worth, cost, expensive, $, reasonable, pricey, cheaper
room	room, suite, view, bed, suite, bathroom, shower, desk, well-equipped, balcony
location	location, traffic, minute, restaurant, locations, mclintock, chandler, located, convenient, mall
cleanliness	clean, dirty, maintain, smell, spotless, tidy, roomy, neat, comfortable, decorated
check in/front desk	stuff, check, help, reservation, check-in, check-outs, flights, appointment, doctor, tech
service	service, food, breakfast, buffet, staff, customer, exceptional, ambiance, friendly, experience
business	business, center, computer, internet, businesses, biz, collier, printer, desktop, wifi
BeerAdvocate	**Keywords**
feel	feel, dryness, softness, sharpness, touch
look	look, appearance, color, dark, transparency
smell	smell, aroma, nose, smelly, snif
taste	taste, flavor, sugary, earthy, bitter

Table 4: Aspect-related keywords

Review	
very unwelcoming staff - downright unfriendly while the room be lovely , the staff be very unfriendly and discourteous . we be very easygoing people . and experienced traveller . however , the staff be very unwilling to answer basic question unk airport unk and restaurant recommendation . one woman behind the desk just seem to be angry all the time . while i love barcelona - this hotel experience be very unk to unk . definitely not a service orient hotel .	
Room	
lovely be love to and very people - and be hotel restaurant definitely be . this unk however i ,	lovely be room , the while the the unfriendly desk staff unwelcoming behind - downright staff very be woman just
Stuff	
unfriendly very very and unfriendly unwelcoming staff be unwilling very very be all to - downright while easygoing just orient	unwelcoming very unfriendly downright not unwilling - discourteous while and staff and a . the unfriendly unk be the orient

Figure 5: Case study. The left blocks contain the words selected by AB-DMSC, the right blocks contain the words selected by D-MILN. We display 20 words with the highest attention weights for each aspect. We manually label the words related with *Room* (in red) and *Stuff* (in green).

Model	TripAdvisor	BeerAdvocate
BERT-asp	**0.8618**	**0.8795**
BERT-doc	0.7512	0.7562
BERT-enc-fix	0.7852	0.7923
BERT-enc-train	0.7540	0.7613

Table 5: Averaged accuracies of BERT-based models

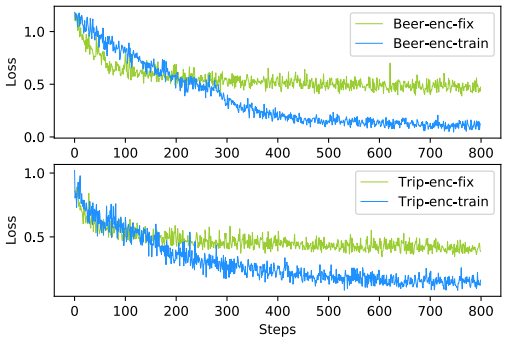


Figure 6: The change of training loss when the training step increases.

BERT-doc, we find that the accuracy declines more than 10% on both datasets when the aspect-level sentiment classifier is trained with document-level annotations with MIL even though the classifier is BERT-based. BERT-enc-fix doesn't outperform D-MILN, we believe this is because the parameters of BERT haven't been fine-tuned for DMSC task. However, when the parameters of BERT are trainable, the performance degrades. By analysing the changes of training loss of BERT-enc-fix and BERT-enc-train, as depicted in Figure 6, we find that the loss of BERT-enc-train declines rapidly to a very low level, showing that it has overfitted the document-level supervision even though the diversified regularizations are applied. In summary, fine-tuning the parameters of BERT with the document-level annotations in MIL will lead to overfitting the document-level sentiment and degrading the performance on aspect-level sentiment prediction. The experiment results also point out a direction for our future work which is to find a way to effectively utilize pre-trained models with weak supervision.

A.3 Case study

To further analyse the performance gap between AB-DMSC and D-MILN, we conduct a qualitative case study on the learned attention mechanism of the aspect-level sentiment classifier. In Figure 5, the gold sentiment labels for *room* and *stuff* are positive and negative respectively. AB-DMSC predicts correctly on both aspects while D-MILN predicts correctly only on *stuff*. For *room*, D-MILN not only picks the words describing it, but also selects the words describing *stuff*. Unfortunately, the words describing *stuff* express an opposite sentiment.

In this case, the description of *room* is much shorter than that of *stuff* and the only words describing *room* are surrounded by the words describing *stuff*. Such unbalanced and mixed descriptions remain a challenge for D-MILN.