Improving Multi-label Emotion Classification via Sentiment Classification with Dual Attention Transfer Network

Jianfei Yu[♣], **Luís Marujo**[♡], **Jing Jiang**[♣], **Pradeep Karuturi**[♡], **William Brendel**[♡]
[♣] School of Information Systems, Singapore Management University, Singapore
[♡] Snap Inc. Research, Venice, California, USA
[♣] jfyu.2014@phdis.smu.edu.sg, jingjiang@smu.edu.sg
[♡] {luis.marujo, pradeep.karuturi, william.brendel}@snap.com

Abstract

In this paper, we target at improving the performance of multi-label emotion classification with the help of sentiment classification. Specifically, we propose a new transfer learning architecture to divide the sentence representation into two different feature spaces, which are expected to respectively capture the general sentiment words and the other important emotion-specific words via a dual attention mechanism. Extensive experimental results demonstrate that our transfer learning approach can outperform several strong baselines and achieve the state-of-the-art performance on two benchmark datasets.

1 Introduction

In recent years, the number of user-generated comments on social media platforms has grown exponentially. In particular, social platforms such as Twitter allow users to easily share their personal opinions, attitudes and emotions about any topic through short posts. Understanding people's emotions expressed in these short posts can facilitate many important downstream applications such as emotional chatbots (Zhou et al., 2018b), personalized recommendations, stock market prediction, policy studies, etc. Therefore, it is crucial to develop effective emotion detection models to automatically identify emotions from these online posts.

In the literature, emotion detection is typically modeled as a supervised multi-label classification problem, because each sentence may contain one or more emotions from a standard emotion set containing *anger*, *anticipation*, *disgust*, *fear*, *joy*, *love*, *optimism*, *pessimism*, *sadness*, *surprise* and *trust*. Table 1 shows three example sentences along with their emotion labels. Traditional approaches to emotion detection include lexicon-based methods (Wang and Pal, 2015), graphical model-based methods (Li et al., 2015b) and linear classifier-based methods (Quan et al., 2015; Li et al., 2015a). Given the recent success of deep learning models, various neural network models and advanced attention mechanisms have been proposed for this task and have achieved highly competitive results on several benchmark datasets (Wang et al., 2016; Abdul-Mageed and Ungar, 2017; Felbo et al., 2017; Baziotis et al., 2018; He and Xia, 2018; Kim et al., 2018).

ID	Tweet	Emotion
T1	AI revolution, soon is possible #fearless #good #goodness	joy, optimism
T2	Shitty is the worst feeling ever #depressed #anxiety	fear, sadness
T3	I am back lol. #revenge	joy, anger

Table 1: Example Tweets from SemEval-18 Task 1.

However, these deep models must overcome a heavy reliance on large amounts of annotated data in order to learn a robust feature representation for multi-label emotion classification. In reality, large-scale datasets are usually not readily available and costly to obtain, partly due to the ambiguity of many informal expressions in user-generated comments. Conversely, it is easier to find datasets (especially in English) associated with another closely related task: sentiment classification, which aims to classify the sentiment polarity of a given piece of text (i.e., positive, negative and neutral). We expect that these resources may allow us to improve sentiment-sensitive representations and thus more accurately identify emotions in social media posts. To achieve these goals, we propose an effective transfer learning (TL) approach in this paper.

Most existing TL methods either 1) assume that both the source and the target tasks share the same sentence representation (Mou et al., 2016) or 2) divide the representation of each sentence into a

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1097–1102
Brussels, Belgium, October 31 - November 4, 2018.

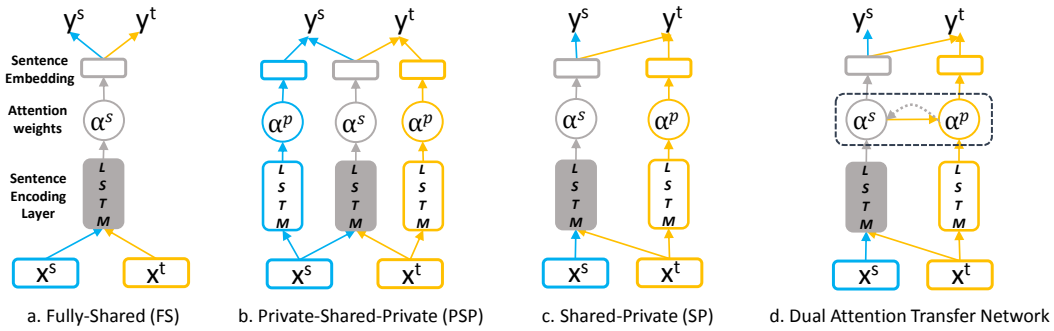

Figure 1: Overview of Different Transfer Learning Models.

shared feature space and two task-specific feature spaces (Liu et al., 2017; Yu et al., 2018), as demonstrated by Fig 1.a and Fig 1.b. However, when applying these TL approaches to our scenario, the former approach may lead the learnt sentence representation to pay more attention to general sentiment words such as *good* but less attention to the other sentiment-ambiguous words like *shock* that are also integral to emotion classification. The latter approach can capture both the sentiment and the emotion-specific words. However, some sentiment words only occur in the source sentiment classification task. These words tend to receive more attention in the source-specific feature space but less attention in the shared feature space, so they will be ignored in our emotion classification task. Intuitively, any sentiment word also indicates emotion and should not be ignored by our emotion classification task.

Therefore, we propose a shared-private (SP) model as shown in Fig 1.c, where we employ a shared LSTM layer to extract shared sentiment features for both sentiment and emotion classification tasks, and a target-specific LSTM layer to extract specific emotion features that are only sensitive to our emotion classification task. However, as pointed out by Liu et al. (2017) and Yu et al. (2018), it is not guaranteed that such a simple model can well differentiate the two feature spaces to extract shared and target-specific features as we expect. Take the sentence **T1** in Table 1 as an example. Both the shared and task-specific layers could assign higher attention weights to *good* and *goodness* due to their high frequencies in the training data but lower attention weights to *fearless* due to its rare occurrences. In this case, this SP model can only predict the *joy* emotion but ignores the *optimism* emotion. Hence, to enforce the orthogonality of the two feature spaces, we further introduce a dual attention mechanism, which feeds the attention weights in one feature space as extra inputs to compute those in the other feature space, and explicitly minimizes the similarity between the two sets of attention weights. Experimental results show that our dual attention transfer architecture can bring consistent performance gains in comparison with several existing transfer learning approaches, achieving the state-of-the-art performance on two benchmark datasets.

2 Methodology

2.1 Base Model for Emotion Classification

Given an input sentence, the goal of emotion analysis is to identify one or multiple emotions contained in it. Formally, let $\mathbf{x} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n)$ be the input sentence with n words, where $\mathbf{w}_j$ is a d-dimensional word vector for word w_j in the vocabulary $\mathcal{V}$, and is retrieved from a lookup table $\mathbf{E} \in \mathbb{R}^{d \times |\mathcal{V}|}$. Moreover, let $\mathcal{E}$ be a set of pre-defined emotion labels. Accordingly, for each $\mathbf{x}$, our task is to predict whether it contains one or more emotions in $\mathcal{E}$. We denote the output as $\mathbf{e} \in \{0,1\}^K$ where $\mathbf{e}_k \in \{0,1\}$ denotes whether or not $\mathbf{x}$ contains the k-th emotion. We further assume that we have a set of labeled sentences, denoted by $D^{\mathrm{e}} = \{\mathbf{x}^{(i)}, \mathbf{e}^{(i)}\}_{i=1}^{N}$.

Sentence Representation: We use the standard bi-directional Long Short Term Memory (Bi-LSTM) network to sequentially process each word in the input:

$$\overrightarrow{\mathbf{h}_j} = \mathrm{LSTM}(\overrightarrow{\mathbf{h}_{j-1}}, \mathbf{x}_j, \Theta_f),$$
$$\overleftarrow{\mathbf{h}_j} = \mathrm{LSTM}(\overleftarrow{\mathbf{h}_{j+1}}, \mathbf{x}_j, \Theta_b),$$

where Θ_f and Θ_b denotes all the parameters in the forward and backward LSTM. Then, for each word $\mathbf{x}_j$, its hidden state $\mathbf{h}_j \in \mathbb{R}^d$ is generated by concatenating $\overrightarrow{\mathbf{h}_j}$ and $\overleftarrow{\mathbf{h}_j}$ as $\mathbf{h}_j = [\overrightarrow{\mathbf{h}_j}; \overleftarrow{\mathbf{h}_j}]$.

For emotion classification, since emotion words are relatively more important for final predic-

tions, we adopt the widely used attention mechanism (Bahdanau et al., 2014) to select the key words for sentence representation. Specifically, we first take the final hidden state $\mathbf{h}_n$ as a sentence summary vector $\mathbf{z}$, and then obtain the attention weight α_i for each hidden state $\mathbf{h}_j$ as follows:

$$u_j = \mathbf{v}^\top \tanh(\mathbf{W}_h \mathbf{h}_j + \mathbf{W}_z \mathbf{z}), \quad (1)$$

$$\alpha_j = \frac{\exp(u_j)}{\sum_{l=1}^{n} \exp(u_l)}, \quad (2)$$

where $\mathbf{W}_h, \mathbf{W}_z \in \mathbb{R}^{a \times d}$ and $\mathbf{v} \in \mathbb{R}^a$ are learnable parameters. The final sentence representation $\mathbf{H}$ is computed as:

$$\mathbf{H} = \sum_{j=1}^{n} \alpha_j \mathbf{h}_j.$$

Output Layer: We first apply a Multilayer Perceptron (MLP) with one hidden layer on top of $\mathbf{H}$, followed by normalizing it to obtain the probability distribution over all of the emotion labels:

$$p(\mathbf{e}^{(i)} \mid \mathbf{H}) = \mathbf{o}^{(i)} = \text{softmax}(\text{MLP}(\mathbf{H})).$$

Then, we propose to minimize the KL divergence between our predicted probability distribution and the normalized ground truth distribution as our objective function:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbf{e}_k^{(i)} \big(\log(\mathbf{e}_k^{(i)}) - \log(\mathbf{o}_k^{(i)}) \big).$$

During the test stage, we will select a threshold γ on the development set so that the emotion with scores higher than γ will be predicted as 1.

2.2 Transfer Learning Architecture

Due to the limited number of annotated data for multi-label emotion classification, here we resort to sentiment classification to consider a transfer learning scenario. Let $D^s = \{\mathbf{x}^{(m)}, y^{(m)}\}_{m=1}^{M}$ be another set of labeled sentences for sentiment classification, where $y^{(m)}$ is the ground-truth label indicating whether the m-th sentence is *positive*, *negative* or *neutral*.

2.2.1 Shared-Private (SP) Model

Intuitively, sentiment classification is a coarse-grained emotion analysis task, and can be fully leveraged to learn a more robust sentiment-sensitive representation. Therefore, we first use a shared attention-based Bi-LSTM layer to transform the input sentences in both tasks into a shared hidden representation $\mathbf{H_c}$, and also employ another task-specific Bi-LSTM layer to get the target-specific hidden representation $\mathbf{H_t}$. Next, we employ the following operations to map the hidden representations to the sentiment label y and the emotion label $\mathbf{e}$:

$$p(y^{(m)}|\mathbf{H_c}) = \text{softmax}\big(\mathbf{W}^s \mathbf{H_c} + \mathbf{b}^s\big),$$
$$p(\mathbf{e}^{(i)}|\mathbf{H_c}, \mathbf{H_t}) = \text{softmax}\big(\text{MLP}([\mathbf{H_c}; \mathbf{H_t}])\big),$$

where $\mathbf{W}^s \in \mathbb{R}^{d \times 3}$ and $\mathbf{b}^s \in \mathbb{R}^3$ are the parameters for the source sentiment classification task.

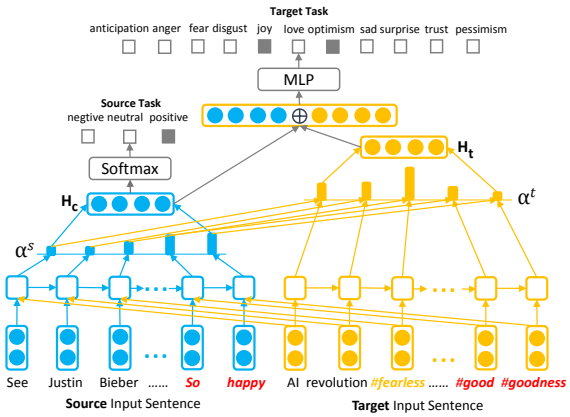

Figure 2: Dual Attention Transfer Network.

2.2.2 Proposed Dual Attention Transfer Network (DATN)

As we introduced before, the shared and target-specific feature spaces in the above SP model are expected to respectively capture the general sentiment words and the task-specific emotion words. However, without any constraints, the two feature spaces may both tend to pay more attention to frequently occurring and important sentiment words like *great* and *happy*, but less to those rarely occurring but crucial emotion words like *anxiety* and *panic*. Therefore, to encourage the two feature spaces to focus on sentiment words and emotion-specific words respectively, we propose using the attention weights computed from the shared layer as extra inputs to compute the attention weights of the target-specific layer. Specifically, as shown in Fig. 2, we first use Eq.1 and Eq.2 to compute the attention weights α^s in the shared layer, and then use the following equation to obtain the attention

weights α^t in the target specific layer:

$$u_j^t = {\mathbf{v}^t}^\top \tanh(\mathbf{W}_h^t \mathbf{h}_j^t + w_\alpha \alpha_j^s + \mathbf{W}_z^t \mathbf{z}^t),$$
$$\alpha_j^t = \frac{\exp(u_j^t)}{\sum_{l=1}^{n} \exp(u_l^t)}.$$

In addition, we introduce another similarity loss to explicitly enforce the difference between the two attention weights and minimize the cosine similarity between α^s and α^t.

Finally, our combined objective function is defined as follows:

$$\mathcal{J} = -\frac{1}{M}\sum_{m=1}^{M} \log p(y_m|\mathbf{H_c}) + \mathcal{L} + \lambda \sum_{i=1}^{N} \text{cos\_sim}(\alpha_i^s, \alpha_i^t),$$

where λ is a hyperparameter used to control the effect of the similarity loss.

2.2.3 Model Details

During the training stage, we adopted the widely used alternating optimization strategy, which iteratively samples one mini-batch from D^{s} for only updating the parameters in the left part of our model, followed by sampling another mini-batch from D^{e} for updating all the parameters in our model. It is also worth noting that in Fig. 2, we first obtain the shared attention weights α^s and feed it as extra inputs to compute α^t. In fact, to differentiate the attention weights in the two feature spaces, we can also first compute α^t, followed by computing α^s based on α^t. We refer to these two variants of our model as DATN-1 and DATN-2 respectively.

3 Experiments

3.1 Experiment Settings

Datasets: We conduct experiments on both English and Chinese languages.

For **English**, we employ a widely used Twitter dataset from SemEval 2016 Task 4A (Nakov et al., 2016) as our source sentiment classification task. For our target emotion classification task, we use the Twitter dataset recently released by SemEval 2018 Task 1C (Mohammad et al., 2018), which contains 11 emotions as shown in the top of Fig. 2. To tokenize the tweets in our dataset, we follow (Owoputi et al., 2013) by adopting most of their preprocessing rules except that we split the hashtag into '#' and its subsequent word.

	Dataset	Train	Dev	Test	Words
E1	SemEval-18	6,838	886	3,259	32,557
S1	SemEval-16	28,631	-	-	40439
E2	Ren-CECps-1	13,841	1,972	3,602	40,099
S2	Ren-CECps-2	15,199	-	-	

Table 2: The number of sentences in each dataset.

For **Chinese**, we use a well known Chinese blog dataset Ren-CECps from (Quan and Ren, 2010), which contains 1487 documents with each sentence labeled by a sentiment label and 8 emotion labels: *anger*, *expectation*, *anxiety*, *joy*, *love*, *hate*, *sorrow* and *surprise*. Given the difficulty of finding a large-scale sentiment classification dataset specific to Chinese blogs, we simply divided the original dataset to form our source and target tasks[1]. The basic statistics of our two datasets are summarized in Table 2.

Parameter Settings: The word embedding size d is set to be 300 for E1 and 200 for E2, and the lookup table $\mathbf{E}$ is initialized by pre-trained word embeddings based on Glove[2]. The hidden dimension and the number of LSTM layers in both datasets are set to be 200 and 1. During training, Adam (Kingma and Ba, 2014) is used to schedule the learning rate, where the initial learning rate is set to be 0.001. Also, the dropout rate is set to 0.5. After tuning, λ is set as 0.05 for both datasets, and γ is set as 0.12 for E1 and 0.2 for E2. All the models are implemented with Tensorflow.

Evaluation Metrics: We take the official code from SemEval-18 Task 1C and use accuracy and Macro F1 score as main metrics. For E2, we follow (Zhou et al., 2018a) to use average precision (AP) and one error (OE) as secondary metrics.

3.2 Results

To better evaluate our proposed methods, we employed the following systems for comparison: 1) *Base*, training our base model in Section 2.1 only on D^{e}; 2) *FT* (Fine-Tuning), using D^{s} to pre-train the whole model, followed by using D^{e} to Fine Tune the model parameters; 3) *FS*, the Fully-Shared framework by (Mou et al., 2016) as shown in Fig 1.a; 4) *PSP* and *APSP*, the Private-Shared-Private framework and its extension with Adver-

[1]The first 560/80/160 documents are used as train/dev/test set for emotion classification, and the remaining 687 documents are used for sentiment classification.

[2]https://nlp.stanford.edu/projects/glove/.

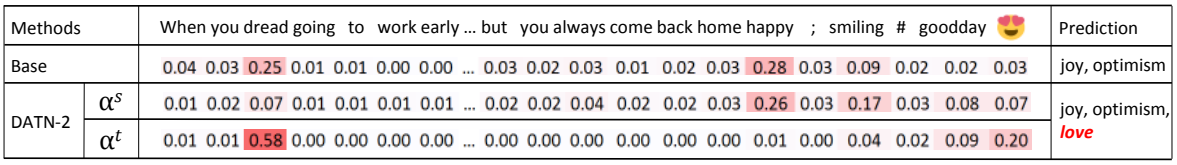

Figure 3: Comparison of attention weights between *Base* and our *DATN-2* model on a test sentence from SemEval-18. Note that the ground truth emotion labels for this example are *joy*, *optimism* and *love*.

Methods	S1 → E1		S2 → E2			
	ACC↑	**F1↑**	**ACC↑**	**F1↑**	**AP↑**	**OE↓**
Base	0.569	0.521	0.368	0.399	0.648	0.531
FT	0.575	0.519	0.372	0.398	0.655	0.519
FS	0.577	0.526	0.386	0.403	0.662	0.507
PSP	0.579	0.531	0.384	0.405	0.658	0.517
APSP	0.580	0.540	0.389	0.399	0.670	0.499
SP	0.577	0.532	0.389	0.410	0.667	0.507
DATN-1	0.582	0.543	0.393	0.410	0.670	0.501
DATN-2	**0.583**	**0.544**	**0.400**	**0.420**	**0.674**	**0.498**
Rank 2	0.582	0.534	-	0.392	0.641	0.523
Rank 1	0.595	0.542	-	0.416	0.680	0.455
DATN-2*	***0.597***	***0.551***	-	-	-	-
Base†	-	-	0.445	0.426	0.725	0.425
DATN-2†	-	-	***0.457***	***0.444***	***0.732***	***0.415***

Table 3: The results of different transfer learning methods by averaging ten runs (top) and the comparison between our best model and the state-of-the-art systems (bottom). DATN-2* indicates the ensemble results of ten runs. Base† and DATN-2† denotes the average results of conducting ten-fold cross validation on the whole dataset for fair comparison, and here for the source and target tasks in DATN-2†, we use the same training data. For E1, Rank1 and Rank2 are the top two systems from the official leadboard; For E2, Rank1 and Rank2 are from (Zhou et al., 2016, 2018a).

sarial losses by (Liu et al., 2017) as shown in Fig 1.b; 5) *SP*, *DATN-1* and *DATN-2*, the Shared-Private model and two variants of our Dual Attention Transfer Network as shown in Fig 1.c and Fig 1.d.

In Table 3, we report the comparison results between our method and the baseline systems. It can be easily observed that 1) for transfer learning, although the performance of *SP* is similar to or even lower than some baseline systems, our proposed dual attention models, i.e., *DATN-1* and *DATN-2*, can generally boost *SP* to achieve the best results. To investigate the significance of the improvements, we combine each model's predictions of all emotion labels followed by treating them as a single label, and then perform McNemar's significance tests (Gillick and Cox, 1989). Finally, we verify that for English, *DATN-1* is significantly better than *Base*, *FT*, *FS* and *SP*, while *DATN-2* is significant better than all the methods except *APSP*; for Chinese, *DATN-1* and *DATN-2* are significantly better than all the compared methods. 2) Even compared with the state-of-the-art systems in E1 which also employ other external resources, including the affective embedding, emotion lexicon and sentiment classification datasets (Baziotis et al., 2018), the ensemble results of *DATN-2* can achieve slightly better performance; in addition, it is clear that our model can obtain the best performance in E2.

Furthermore, to obtain a better understanding of the advantage of our method, we choose one sentence from the test set of E1, and visualize the attention weights obtained by *Base* and *DATN-2* in Fig 3. We can see that *Base* pays more attention to those frequent emotion words while ignoring the less frequent but important emoji, and thus fails to predict the *love* emotion implied by the emoji. In contrast, with the proposed dual attention mechanism, *DATN-2* makes correct predictions since it can respectively capture the general sentiment words and the emotion-specific emojis.

4 Conclusion

In this paper, we proposed a dual attention-based transfer learning approach to leverage sentiment classification to improve the performance of multi-label emotion classification. Using two benchmark datasets, we show the effectiveness of the proposed transfer learning method.

Acknowledgments

We would like to thank Aletta Hiemstra, Andrés Monroy-Hernández and three anonymous reviewers for their valuable comments.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly

learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Christos Baziotis, Nikos Athanasiou, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. 2018. NTUA-SLP at SemEval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning. *arXiv preprint arXiv:1804.06658*.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Laurence Gillick and Stephen J Cox. 1989. Some statistical issues in the comparison of speech recognition algorithms. In *International Conference onAcoustics, Speech, and Signal Processing*.

Huihui He and Rui Xia. 2018. Joint binary neural network for multi-label learning with applications to emotion classification. *arXiv preprint arXiv:1802.00891*.

Yanghoon Kim, Hwanhee Lee, and Kyomin Jung. 2018. Attnconvnet at semeval-2018 task 1: Attention-based convolutional neural networks for multi-label emotion classification. *arXiv preprint arXiv:1804.00831*.

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

Li Li, Houfeng Wang, Xu Sun, Baobao Chang, Shi Zhao, and Lei Sha. 2015a. Multi-label text categorization with joint learning predictions-as-features method. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Shoushan Li, Lei Huang, Rong Wang, and Guodong Zhou. 2015b. Sentence-level emotion classification with label and context dependence. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Saif M Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *SemEval*.

Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How transferable are neural networks in NLP applications? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. Semeval-2016 task 4: Sentiment analysis in twitter. In *SemEval*.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL*.

Changqin Quan and Fuji Ren. 2010. Sentence emotion analysis and recognition based on emotion words using ren-cecps. *International Journal of Advanced Intelligence*.

Xiaojun Quan, Qifan Wang, Ying Zhang, Luo Si, and Liu Wenyin. 2015. Latent discriminative models for social emotion detection with emotional dependency. *ACM Transactions on Information Systems (TOIS)*, 34(1):2.

Yaqi Wang, Shi Feng, Daling Wang, Ge Yu, and Yifei Zhang. 2016. Multi-label chinese microblog emotion classification via convolutional neural network. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data*.

Yichen Wang and Aditya Pal. 2015. Detecting emotions in social media: A constrained optimization approach. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Jianfei Yu, Minghui Qiu, Jing Jiang, Jun Huang, Shuangyong Song, Wei Chu, and Haiqing Chen. 2018. Modelling domain relationships for transfer learning on retrieval-based question answering systems in E-commerce. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*.

Deyu Zhou, Yang Yang, and He Yulan. 2018a. Relevant emotion ranking from text constrained with emotion relationships. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Deyu Zhou, Xuan Zhang, Yin Zhou, Quan Zhao, and Xin Geng. 2016. Emotion distribution learning from texts. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018b. Emotional chatting machine: emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*.