Recognizing Conflict Opinions in Aspect-level Sentiment Classification with Dual Attention Networks

Xingwei Tan, Yi Cai[*] and **Changxi Zhu**
School of Software Engineering, South China University of Technology
Guangzhou, China
setanxw5@mail.scut.edu.cn
*ycai@scut.edu.cn

Abstract

Aspect-level sentiment classification, which is a fine-grained sentiment analysis task, has received lots of attention these years. There is a phenomenon that people express both positive and negative sentiments towards an aspect at the same time. Such opinions with conflicting sentiments, however, are ignored by existing studies, which design models based on the absence of them. We argue that the exclusion of conflict opinions is problematic, for the reason that it represents an important style of human thinking – dialectic thinking. If a real-world sentiment classification system ignores the existence of conflict opinions when it is designed, it will incorrectly mixed conflict opinions into other sentiment polarity categories in action. Existing models have problems when recognizing conflicting opinions, such as data sparsity. In this paper, we propose a multi-label classification model with dual attention mechanism to address these problems.

1 Introduction

Aspect-level sentiment classification is a fine-grained sentiment analysis task (Pang et al., 2008). Given aspect categories or target entities in the text, this task aims at inferring the sentiment polarity (e.g., *positive*, *negative*, or *neutral*) of the aspects. A recent study (Kenyon-Dean et al., 2018) on sentiment analysis dataset construction brings attention to sentiments beyond the scope of *positive*, *negative*, and *neutral* categories, indicating that they will benefit the real-world sentiment analysis systems. We find that there is a similar type of sentiment which is labeled as *conflict* in SemEval dataset for aspect-level sentiment classification (Pontiki et al., 2014). *Conflict* label applies when both positive and negative sentiment is expressed about an aspect. For example, the

sentence "*The atmosphere is attractive, but a little uncomfortable.*" expresses *conflict* sentiment towards *ambience* aspect. However, most of existing studies ignore *conflict* opinions, for the reason that they are sparse in the datasets (Tang et al., 2016b; He et al., 2018). In this work, we argue that the exclusion of *conflict* opinions is problematic.

The *conflict* opinions are the production of an important style of human thinking – dialectic thinking. Dialectic thinking, unlike the thinking constrained by the laws of formal logic, shows a degree of tolerance of contradiction (Peng and Nisbett, 1999). If a real-world sentiment classification system ignores the existence of *conflict* opinions when it is designed, its predicted *positive* opinions may contain negative expressions and its predicted *negative* opinions may contain positive expressions. Considering the existence of sentiment expressions of both sites, it is unreasonable to put *conflict* opinions into either *positive* or *negative* sentiment classes.

There are problems when including *conflict* opinions in aspect-level sentiment classification. Based on our observation, existing methods have significant lower accuracy on recognizing *conflict* sentiment than recognizing other sentiments. We analyse that the reasons are as follows: 1) *Conflict* opinions are sparse in dataset, which is difficult for a model to learn its class-specific features during training; 2) Existing models have difficulty in recognizing the complex expressions of *conflict* sentiment, which are the combination of positive and negative sentiment expressions.

In this paper, we propose a method to address the two problems. Firstly, we model the task as a multi-label classification problem. In contrast to mutually exclusive 4-class (*positive*, *negative*, *neutral*, and *conflict*) classification, our model predicts two labels for each aspect: 1) Whether positive sentiment is expressed towards the aspect; 2)

---

[*]Corresponding author

Whether negative sentiment is expressed towards the aspect. Then, we transform the 2-label predicted results into 4-class labels by rules (e.g., if both positive and negative expressions exist, then apply *conflict*). During training, the 4-class training labels in dataset are transformed into 2-label targets. In this way, we utilize the relation between *conflict* label and *positive*/*negative* labels, which makes the model learn to recognize *conflict* opinions from abundant *positive* and *negative* data. It is also based on the fact that neural network is good at recognizing positive or negative sentiments, which are indicated by compact and concrete expressions. On the contrary, we find that the expressions of *conflict* opinions usually are lengthy and implicit. Faced with *conflict* opinions, existing attention-based models (Wang et al., 2016) (with single attention) usually attend on a portion of a *conflict* expression and ignore the other portion, which make its predicted result prone to be *positive* or *negative*. Instead, we use two independent attentions in the proposed model to focus on positive or negative expressions separately.

Our contributions are summarized as follows: 1) To the best of our knowledge, this is the first work that discuss the necessity of including *conflict* opinions in aspect-level sentiment classification task; 2) We propose a model that can recognize the complex expressions of *conflict* opinions and deal with sparsity problem of *conflict* data; 3) The experiment results on SemEval dataset indicate that our model performs well on recognizing *conflict* opinions and outperforms all baselines.

2 Related Work

Sentiment analysis is a frequently studied topic in artificial intelligence (Cambria, 2016). Different from previous work about sentiment analysis, Kenyon-Dean et al. (2018) states that there are opinions cannot be properly annotated as *positive*, *negative*, or *neutral* sentiment polarities. Their experiment shows that existing classifiers have significant lower accuracy on recognizing sentiments in these opinions. Their work is about sentence-level sentiment analysis and does not propose a model. Our work is different from theirs since we propose a model capable of recognizing *conflict* opinions in existing aspect-level sentiment classification dataset.

In recent years, lots of efforts have been made

to imporve aspect-based sentiment analysis (Wang et al., 2014; Zhao et al., 2015). Traditional methods in aspect-level sentiment classification leveraged sentiment lexicons (Mohammad et al., 2013), while recent work tends to use neural network to generate text representations automatically. TD-LSTM and TC-LSTM (Tang et al., 2016a) encode contextual information with respect to the target. Methods based on memory network (Tang et al., 2016b; Zhu and Qian, 2018) instead generate aspect-related representation. Attention-based recurrent neural network (Wang et al., 2016; Chen et al., 2017; Cheng et al., 2017) selectively attends to aspect-related regions and calculates the weighted sum of the hidden states of these regions. Xue et al. (2018) instead utilize gated convolutional neural network to extract aspect-related representations, which increase computational efficiency. Our work differs from these works since we address two problems of recognizing *conflict* opinion which they ignore.

3 Model: Dual Attention-based GRU

3.1 Network Architecture

The architecture of the proposed Dual Attenition-based GRU (D-AT-GRU) is illustrated as Fig. 1. Given a sentence with L words and an aspect (or aspect term) as inputs, the model first transforms them into word embeddings $[w_1, ..., w_L]$ and aspect embedding v_a. Then, it extracts aspect-related text features through recurrent neural network and dual attention mechanism. Lastly, it produces 2-label predictions through classification layers and transforms them into 4-class labels.

Aspect embedding. We use an embedding matrix $V_A \in \mathbb{R}^{|A| \times d_a}$ to represent aspects, where d_a is the dimension of aspect embedding and $|A|$ is the number of aspects in dataset. Each row in V_A is a representation vector v_a for aspect $a \in A$. The aspect embeddings are initialized randomly and learned as parameters.

Text feature extraction. We use Gated Recurrent Unit (GRU) (Cho et al., 2014) to extract contextual information of words in our model. Long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) is an alternative choice, which is often used in previous models (Wang et al., 2016; Chen et al., 2017; He et al., 2018). However, we find that GRU performs better than LSTM in our model. Thus, we choose GRU. Let matrix $H = [h_1, ..., h_L] \in \mathbb{R}^{L \times d}$ denotes the hid-
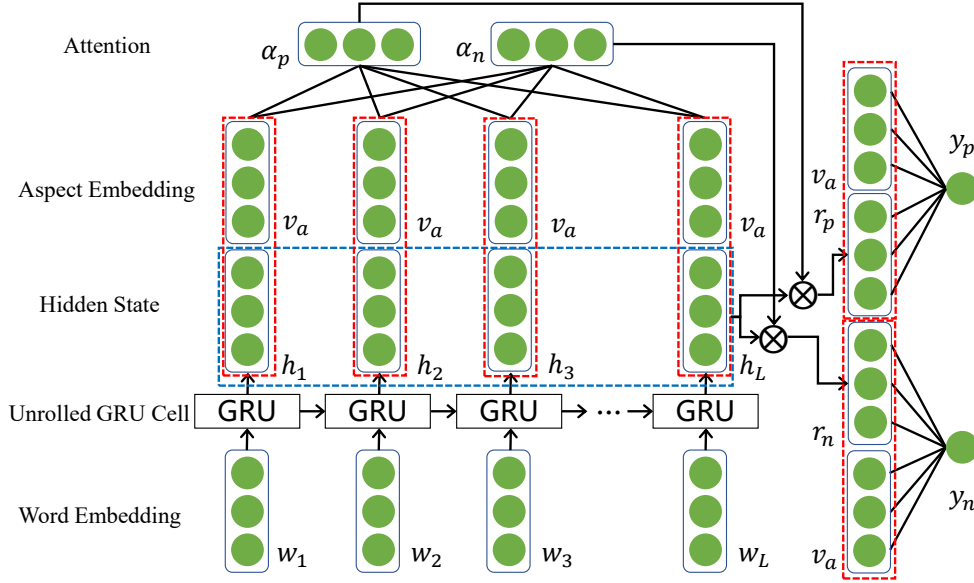
Figure 1: The network architecture of D-AT-GRU.

den states produced by GRU.

Attention mechanism. Based on the hidden states H and the embedding v_a of given aspect, attention mechanism gives importance weights which indicate aspect-related context region. Given that a *conflict* opinion consist of positive and negative expressions, attention usually fails to capture the full expression and only attends to a portion of it. Affected by previous *positive* or *negative* training samples, attention tends to think that just a portion of the *conflict* expression is enough to determine sentiment label and omit the other portion which also contains sentiment words. We therefore propose to use two attentions separately: one is responsible for finding positive expression, the other is responsible for finding negative expression. The positive attention is calculated by:

$$g_i^p = u_p^T tanh(W_p[h_i; v_a] + b_p), \qquad (1)$$

$$\alpha_i^p = \frac{exp(g_i^p)}{\sum_{j=1}^L exp(g_j^p)}, \qquad (2)$$

where $W_p \in \mathbb{R}^{d \times 2d}, u_p \in \mathbb{R}^d$ and $b_p \in \mathbb{R}^d$ are parameters. The negative attention is calculated through following functions:

$$g_i^n = u_n^T tanh(W_n[h_i; v_a] + b_n), \qquad (3)$$

$$\alpha_i^n = \frac{exp(g_i^n)}{\sum_{j=1}^L exp(g_j^n)}, \qquad (4)$$

where $W_n \in \mathbb{R}^{d \times 2d}, u_n \in \mathbb{R}^d$ and $b_n \in \mathbb{R}^d$ are parameters.

Regularization term. For the reason that positive sentiment and negative sentiment usually are expressed in different regions, the two attentions should not overlap. Thus, we need to add a regularization term to make the two sets of attention scores differentiate from each other. We compare KL-divergence (Kullback and Leibler, 1951) and orthogonal regularization (He et al., 2017) in experiments. We observed no consistent difference in performance regarding to sentiment analysis between these two terms. But the time cost of computing KL-divergence is 2.7 times as much as computing orthogonal regularization for each batch. Therefore, we utilize orthogonal regularization (He et al., 2017):

$$\mathcal{R}_o = ||M^T M - I||_2, \qquad (5)$$

where $M = [\alpha^p; \alpha^n], M \in \mathbb{R}^{2 \times L}$ contains attention scores.

Text representation. According to the positive and negative attention scores, corresponding aspect-related text representations are calculated through the weighted sum of hidden states H:

$$v_p = \sum_{i=1}^L \alpha_i^p h_i, \qquad (6)$$

$$v_n = \sum_{i=1}^L \alpha_i^n h_i. \qquad (7)$$

Classification layer. An expression may have various meanings when describing different aspects. Therefore, we take aspect information into

3428

account in classification layer by concatenating text representations with aspect embeddings.

$$y_p = \sigma(W_{py}[v_p; v_a] + b_{py}), \qquad (8)$$
$$y_n = \sigma(W_{ny}[v_n; v_a] + b_{ny}), \qquad (9)$$

where y_p is predicted positive sentiment distribution, y_n is predicted negative sentiment distribution; $W_{py}, W_{ny} \in \mathbb{R}^{2d}$ and $b_{py}, b_{ny} \in \mathbb{R}^1$ are parameters.

3.2 Objection Function

The model can be trained in an end-to-end way by backpropagation, where the objective function is the sum of the binary cross-entropy losses of positive and negative sentiments:

$$\mathcal{L}_p = -\hat{y}_p log(y_p) - (1 - \hat{y}_p)log(1 - y_p), \quad (10)$$
$$\mathcal{L}_n = -\hat{y}_n log(y_n) - (1 - \hat{y}_n)log(1 - y_n), \quad (11)$$

where $\hat{y}_p, \hat{y}_n$ are target distributions. During training, 4-class sentiment labels are transformed into 2-label targets (e.g., *conflict* is transformed into $(\hat{y}_p, \hat{y}_n) = (1, 1)$). The goal of training is to minimize the following function:

$$\mathcal{L} = \mathcal{L}_p + \mathcal{L}_n + \lambda \mathcal{R}_o, \qquad (12)$$

where λ is a hyperparameter used to balance the weight of orthogonal regularization. We set λ to 1 in our experiment.

3.3 Label Transformation

Given the predicted distributions y_p, y_n, we transform them into 4-class labels. If $y_p > p, y_n > p$, the predicted label will be *conflict*. If $y_p \leq p, y_n \leq p$, the predicted label will be *neutral*. If $y_p \leq p, y_n > p$, the predicted label will be *negative*. If $y_p > p, y_n \leq p$, the predicted label will be *positive*. In this paper, we directly use $p = 0.5$ to minimize human interference, although it can be tuned manually as a hyperparameter.

4 Experiment

4.1 Dataset and Preparation

We experiment on SemEval 2014 datasets (Pontiki et al., 2014) (statistics shown in Table 1) to evaluate the proposed model. For the reason that there are too few *conflict* test instances regarding aspect term to show significance, we choose to experiment on aspect category data. Our model is compared with the state-of-the-art methods in aspect category sentiment classification, which are adapted to 4-class classification.

In our experiment[1], dimension sizes d, d_a, d_w are all set to 300. Word embeddings are initialized with 300-dimension pre-trained GloVe vectors (Pennington et al., 2014). All the other parameters are initialized by sampling from a normal distribution $\mathcal{N}(0, 0.01)$. We use Adagrad (Duchi et al., 2011) with a batch size of 10 samples, initial learning rate of 0.01, and maximum epochs of 30. We use development set chosen by Tay et al. (2018). The patience value of early-stopping is 3.

	positive	*negative*	*neutral*	*conflict*
Train	1759	704	442	187
Dev	420	135	58	8
Test	657	222	94	52

Table 1: Statistics of datasets.

4.2 Result and Analysis

Model	Overall	Conflict
AT-LSTM	77.13%	11.54%
ATAE-LSTM	78.00%	23.08%
GCAE	78.30%	25.00%
AT-GRU	77.22%	19.23%
AT-GRU 2-label	77.02%	26.92%
D-AT-GRU w/o orthogonal	77.22%	26.92%
D-AT-GRU	**78.50%**	**40.38%**

Table 2: Experiment results on SemEval 2014 dataset.

Table 2 shows that D-AT-GRU model outperforms all baseline methods. Given that AT-LSTM (Wang et al., 2016) has strong correlation to our base model (AT-GRU), their work can be categorized as a baseline to our model. In contrast, the results prove that the additional components are helpful to recognize *conflict* opinions. We also compare our model to the recently proposed GCAE (Xue and Li, 2018), which is based on gated CNN. D-AT-GRU performs competitively with GCAE overall and significantly better on *conflict* category.

We also compare with ablated versions of D-AT-GRU: 1) Replace the LSTM in AT-LSTM with
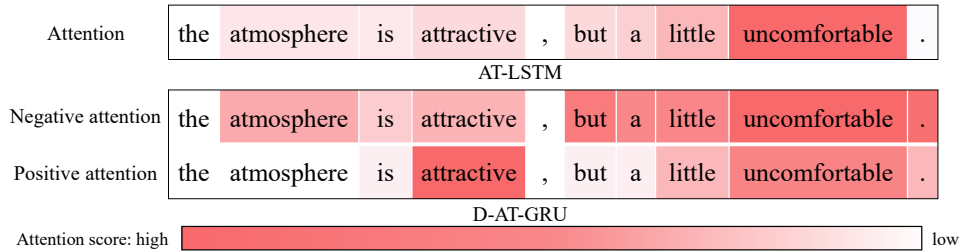

Figure 2: The visualizations of attention.

GRU and directly do 4-class classification (AT-GRU); 2) Use AT-GRU with two separate classification layers to do 2-label classification (AT-GRU 2-label); 3) Remove orthogonal regularization (D-AT-GRU w/o orthogonal). Their performance suffers from the ablations, which proves that every component is essential for the proposed model.

We test KL-divergence (Kullback and Leibler, 1951) and orthogonal regularization (He et al., 2017) as regularization terms to make positive attention and negative attention differentiate from each other. The experiments were conducted on a NVIDIA 1080 Ti. The results are shown in Table 3.

Regularizer	Overall	Conflict	Time
Orthogonal	78.50%	40.38%	0.15 ms
KL-divergence	78.40%	40.38%	0.40 ms

Table 3: Experiment results about different regularization terms.

4.3 Case Study

Figure 2 is a test case in our experiment. AT-LSTM attends to "*uncomfortable*" and ignores the existence of "*attractive*", which causes it to incorrectly predict *negative* label. With our D-AT-GRU model, the negative attention finds "*uncomfortable*" in the second part of the sentence, which makes the model predict that there is negative sentiment; the positive attention instead gives the highest score to "*attractive*", which leads the model to predict that there is positive sentiment. Given the existence of both positive and negative sentiments, D-AT-GRU correctly predict *conflict* label. This shows that the dual attention mechanism of D-AT-GRU model can more accurately capture the complete expression of a *conflict* opinion.

5 Conclusion

In this paper, we present a method to recognize *conflict* opinions in aspect-level sentiment classification task. By transforming the problem into recognizing simpler sentiments, we alleviate the sparsity problem of *conflict* data. Our model utilize dual attention mechanism and orthogonal regularization, which are capable of recognizing the complex expressions of *conflict* opinions. Our experiment on SemEval dataset demonstrates the effectiveness of the proposed model.

References

Erik Cambria. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31:102–107.

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.

Jiajun Cheng, Shenglin Zhao, Jiani Zhang, Irwin King, Xin Zhang, and Hui Wang. 2017. Aspect-level sentiment classification with heat (hierarchical attention) network. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 97–106. ACM.

Kyunghyun Cho, Bart van Merrienboer, aglar Gülehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *ACL*.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2018. Effective attention modeling for aspect-level sentiment classification. In *COLING*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, et al. 2018. Sentiment analysis: Its complicated! 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 1886–1895.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Saif Mohammad, Svetlana Kiritchenko, and Xiao-Dan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *SemEval@NAACL-HLT*.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Kaiping Peng and Richard E Nisbett. 1999. Culture, dialectics, and reasoning about contradiction. *American psychologist*, 54(9):741.

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.

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

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016a. Effective lstms for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307.

Duyu Tang, Bing Qin, and Ting Liu. 2016b. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224.

Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. 2018. Learning to attend via word-aspect associative fusion for aspect-based sentiment analysis. In *AAAI*.

Tao Wang, Yi Cai, Ho fung Leung, Raymond Y. K. Lau, Qing Li, and Huaqing Min. 2014. Product aspect extraction supervised with online domain knowledge. *Knowl.-Based Syst.*, 71:86–100.

Yequan Wang, Minlie Huang, Li Zhao, et al. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

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)*, volume 1, pages 2514–2523.

Y. Zhao, B. Qin, and T. Liu. 2015. Creating a fine-grained corpus for chinese sentiment analysis. *IEEE Intelligent Systems*, 30(1):36–43.

Peisong Zhu and Tieyun Qian. 2018. Enhanced aspect level sentiment classification with auxiliary memory. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1077–1087.