Aspect-based Sentiment Analysis with LSTM and BERT

Xinli Gu
NYU CDS
xg588
xg588@nyu.edu

Di He
NYU CDS
dh3171
dh3171@nyu.edu

Yuchuan Fu
NYU CDS
yf2127
yf2127@nyu.edu

Lining Zhang
NYU CDS
lz2332
lz2332@nyu.edu

Abstract

Aspect-based sentiment analysis (ABSA) is a fine-grained textual classification task, which predicts the sentiment polarity of a sentence given a certain aspect. In this paper, we implement LSTM-based models (LSTM and ATAE-LSTM) and BERT-based models (vanilla BERT-base and BERT-ADA) on a challenging dataset called MAMS, which contains multiple aspects with multiple sentiment polarities. We also conduct error analysis through the method of input reduction. The experiment results show that BERT-ADA outperforms other models on MAMS dataset.

1 Introduction

Sentiment Analysis (SA) aims at classifying the sentiment polarity towards a whole sentence. Compared to SA, Aspect-Based Sentiment Analysis (ABSA) is designed to identify certain target aspects of an entity and classify sentiment polarities towards these aspects. For example, in the sentence "Food is pretty good but the service is horrific", there are two target aspects: "food" and "service" with opposite sentiment polarities. Further, there are two variants of the ABSA problem. One is Aspect-Target Sentiment Classification (ATSC), which is our example before and also the focus of our paper.

In recent years, neutral networks have been developed and largely improved the ABSA performance by learning target-context relationships. Afterwards, the pre-trained language model shows powerful representation ability. Its application to many down-stream tasks, including ABSA, has achieved many accomplishments. Lately, domain-specific post-trained BERT shows better performance on this topic.

In this paper, we experiment with LSTM-based and BERT-based models for aspect-based sentiment analysis, and apply these models to a more challenging dataset than the commonly used benchmark. We then conduct a robust error analysis for our models to analyze reasons for erroneous classifications.

2 Related Work

2.1 Traditional Aspect-based Sentiment Analysis

Aspect-based Sentiment Analysis (ABSA) is a fine-grained textual classification task, which predicts the sentiment polarity of a sentence given a certain aspect. Previously, this task heavily relied on manually-designed lexicon-based features. These early works focus on sentiment classification with features, like bag-of-words and sentiment lexicons (Rao and Ravichandran, 2009). Specific methods include SVM (Mullen and Collier, 2004), rule-based methods (Ding et al., 2008), and statistic-based methods (Jiang et al., 2011). However, these traditional aspect-based sentiment classifiers rely on high-quality feature-engineering which is labor intensive.

2.2 Aspect-based Sentiment Analysis with Neural Networks

In recent years, deep neural networks have achieved great success in the task of Aspect-based Sentiment Analysis (ABSA) through automatic learning of textual representation (Dong et al., 2014; Nguyen and Shirai, 2015). Among these works, ATAE-LSTM combines LSTM architecture with attention mechanism and includes embedding vectors of aspects to participate in computing attention weights (Wang et al., 2016). The convolutional neural networks have also been applied in some of the model architectures with slight modification for modeling natural language tasks (Huang and Carley, 2018).

With the advance of model architecture, transformer (Vaswani et al., 2017) and BERT-based

methods (Devlin et al., 2018) have performed well on the ABSA task. In some works, the pre-trained BERT architecture weights are finetuned on a domain-specific corpus and trained for the ABSA task (Rietzler et al., 2019). Furthermore, an interactive multi-task learning network (IMN) has also been proposed for end-to-end aspect-based sentiment analysis (He et al., 2019).

However, most of these works are experimented on the SemEval dataset (Pontiki et al., 2014), which consists of only one aspect or multiple aspects with the same sentiment polarity. Thus, we conduct our experiment through LSTM and BERT based methods on a dataset called Multi-Aspect Multi-Sentiment (MAMS) dataset (Jiang et al., 2019), which makes the ABSA task more challenging.

3 Models

3.1 Long Short-term Memory(LSTM)

Recurrent Neural Network(RNN) can perform various NLP tasks due to its merit of handling sequential data. However, RNN suffers from the vanishing or exploding gradient problem which prevent RNN from capturing longer dependency. Therefore, Long Short-term Memory(LSTM) was developed in order to concur this problem. In addition from vanilla RNN model, LSTM has three gates(forget, input and output) and a cell state. Specifically:

$$X = \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} \tag{1}$$

$$f_t = \sigma(W_f \cdot X + b_f) \tag{2}$$

$$i_t = \sigma(W_i \cdot X + b_i) \tag{3}$$

$$o_t = \sigma(W_o \cdot X + b_o) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c \cdot X + b_c) \tag{5}$$

$$h_t = o_t \odot \tanh(c_t) \tag{6}$$

where each hidden layer h_t has the same dimension d as the input word embedding x_t. W_i, W_f, $W_o \in \mathbb{R}^{d \times 2d}$ are the weight matrices and b_i, b_f, $b_o \in \mathbb{R}^d$ are biases for the input, forget, and output gate respectively. σ is the sigmoid function and $\odot$ represents element-wise multiplication. The last timestamp hidden state h_N is the representation of the whole sentence and we will pass h_N to a linear layer which will project the hidden state to number of classes, which is 3 in our case.

3.2 Attention-based LSTM with Aspect Embedding (ATAE-LSTM)

The previous method of LSTM doesn't take aspect into consideration, and thus the result will be more similar to sentence level sentiment classification. We would expect it to perform poorly on the MAMS dataset which includes multiple aspects and multiple sentiments. In order to better capture the aspect information, we adopt the idea from Wang et al. (2016), who proposed to append aspect embedding to each input vector and include the attention mechanism that can detect the important parts in a sentence regarding the specific aspect. Figure 1 represents the architecture of an Attention-based LSTM with Aspect embedding(ATAE-LSTM).

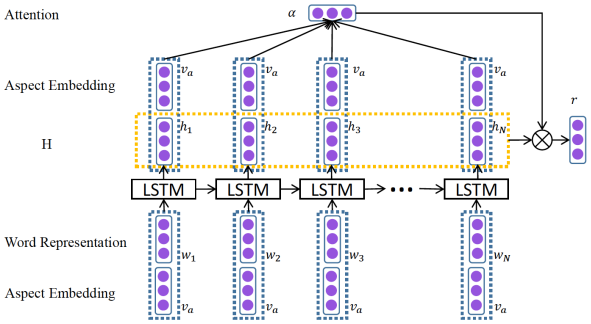

Figure 1: The Architecture of Attention-based LSTM with Aspect Embedding

With aspect embeddings, the output hidden states $(h_1, h_2, \ldots, h_N)$ can have information from the current aspect (v_a). With this modification, for every time step t, the input vector becomes:

$$input_t = \begin{bmatrix} x_t \\ v_a \end{bmatrix} \tag{7}$$

where $v_a \in \mathbb{R}^d_a$ represents the embedding of aspect a.

Let $H \in \mathbb{R}^{d \times N}$ be the hidden matrix where d is the hidden dimension and N is the sequence length. $e_N \in \mathbb{R}^N$ is a vector of 1s. The attention mechanism will produce an attention weight vector α and a weighted hidden representation r as follows.

$$M = \tanh(\begin{bmatrix} W_h H \\ W_v v_a \bigotimes e_N \end{bmatrix}) \tag{8}$$

$$\alpha = softmax(\omega^T M) \tag{9}$$

$$r = H\alpha^T \tag{10}$$

where $M \in \mathbb{R}^{(d+d_a)\times N}$, $\alpha \in \mathbb{R}^{d}$, $W_h \in \mathbb{R}^{d\times d}$, $W_v \in \mathbb{R}^{d\times d_a}$, $\omega \in \mathbb{R}^{d+d_a}$ and $v_a \bigotimes e_N = [v; v; \ldots; v]$. Then we can compute the new hidden layer by combining r and h_N. Finally, we project the new hidden layer to dimension of number of classes and apply a softmax layer to it to obtain the probability representation of the sentence with respect to current aspect.

3.3 Bidirectional Encoder Representations from Transformers (BERT)

The BERT model, based on many previous innovations, has a deeply bidirectional architecture and could create powerful representations. The pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models on many downstream tasks (Devlin et al., 2018). Based on the same way that the above paper proposed for sequence-pair classification tasks, we use a tokenized sentence together with a target term as input.

As Figure 2 shows, we transform the reviews data into "[CLS] sentence [SEP] target term [SEP]" form as the input. If one sentence has multiple target terms, it will be treated as input multiple times, every time with a single target term. Because we use the BERT-base model, the last hidden representation of the [CLS] token will be $\in \mathbb{R}^{786\times 1}$ and the notation is $h_{[CLS]}$. The output is three sentiment polarity classes: positive, natural and negative. We get the output from a linear transformation followed by a softmax activation function:

$$p = softmax(W \cdot h_{[CLS]}) + b$$

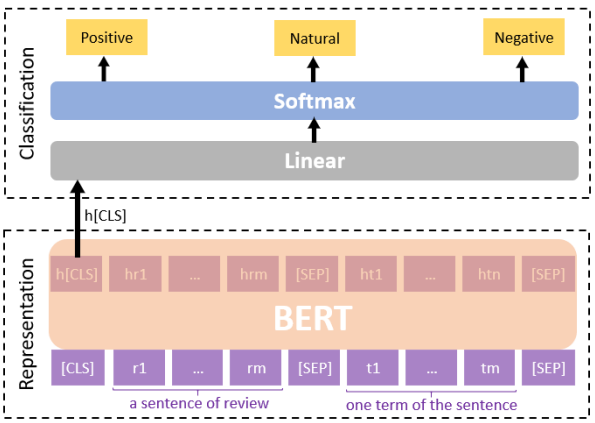

Figure 2: Overall procedures for ABSC with BERT

3.4 BERT with Domain Adaptation (BERT-ADA)

However, finetuning BERT directly on the end task based on limited tuning data could have domain and task awareness challenges. Based on masked language model (MLM) and next-sentence prediction (NSP), post-tuned domain-specific language model could alleviate both problems (Xu et al., 2019). The BERT-ADA which is post-trained on a restaurant corpus achieves a new state-of-the-art performance on the SemEval 2014 restaurants dataset (Rietzler et al., 2019). Therefore, we use its BERT-ADA model to investigate the performance of domain-specific BERT model on multi-aspect multi-sentiment classification problems. The procedures of implementing BERT-ADA are the same as BERT-base, as illustrated in Figure 2.

3.5 Input Reduction for Model Interpretation

Feng (Feng et al., 2018) introduced the input reduction method for model interpretation. The goal of this method is to find a subset of the most important words that contribute to a prediction. We use this method to investigate the reasons of erroneous classifications. Samples will be drawn from erroneous predicted test data. By removing input words from those samples iteratively, words that change predicted polarities will be marked as our keywords. Afterwards, a deeper analysis will be conducted.

4 Experiments

4.1 Dataset and experimental settings

4.1.1 Dataset Overview

We conduct experiments using the SemEval 2014 Task 4 Subtask 2 restaurant dataset, and the MAMS restaurant dataset. Table 1 shows the overview of the two datasets. In this project, We focus on the aspect-term sentiment analysis only. Each sentence in the MAMS dataset contains at least two terms, and at least two aspects in the same sentence have different emotional polarities. The MAMS dataset has a 3-level sentiment polarity (positive, negative

Dataset		Pos.	Neg.	Neu.	Total.
SemEval-14	Train	2164	805	724	3693
	Test	728	196	210	1134
MAMS	Train	3783	3089	5646	12518
	Test	400	329	607	1336

Table 1: Overview of SemEval and MAMS Dataset.

or neutral), while the SemEval has an additional label 'conflict'. During our experiment, the conflict labels are dropped for reasons of comparability.

4.1.2 Implementation Details

For all the non-BERT based models, we use 300-dimentional word vectors pre-trained by GloVe to initialize the word embedding vectors. For LSTM and ATAE-LSTM, we use Adam optimizer with a learning rate of 2e-5. We train these two models with a batch size of 16, and L2-regularization weight of 0.01. For the BERT-based models, we first load the BERT-base and BERT-ADA models and then use them for classification. We train the models with a batch size of 32. Adam optimizer is also used but with a learning rate of 3e-5.

4.2 Comparative models

We adopt several state-of-the-art as well as baseline models during our experiment, which we will now describe briefly.

LSTM: Traditional LSTM cannot capture any information about the aspects in the sentence, therefore would have the worst performance among all.

ATAE-LSTM: ATAE-LSTM first attaches the aspect embedding to each word embedding, and then employs attention mechanism to get the sentence representation for final classification. It can capture the important and different parts of a sentence when given different aspects.

CapsNet: The CapsNet model (Jiang et al., 2019) consists of an embedding layer, an encoding layer, a primary capsule layer and a category capsule layer.

CapsNet-BERT: CapsNet-BERT combines capsule network with BERT-base, which replaces the embedding layer and encoding layer of CapsNet with pre-trained BERT.

BERT-base: BERT-base is using the pretrained BERT-base embeddings directly on the downstream task without any domain specific language model finetuning.

BERT-ADA: BERT-ADA is the BERT model that has been finetuned on restaurant domain corpora, the Yelp Dataset Challenge reviews (https://www.yelp.com/dataset/challenge).

4.3 Result analysis

Experiment results (Accuracy and F1-score) are reported in Table 1. First, all models perform better on the SemEval-14 Restaurant Review dataset than MAMS dataset, verifying that the MAMS dataset is more challenging. The reason may be that for MAMS, every sentence has at least two labels, and two opposite polarities. The labels conflict with each other when each sentence is extended to several samples. For GloVe based models, traditional LSTM cannot capture any aspect information, therefore it's the worst. However, with Attention-based LSTM, there's a huge improvement especially on the MAMS dataset. For BERT-based models, they also generalize well on MAMS.

Second, the domain-trained BERT-ADA outperforms all other models and achieves the best performance. Compared with vanilla BERT-base and CapsNet-BERT, pretraining on specific target domain indeed produces a huge improvement.

4.4 Error Analysis: Case Study

In this part, we investigate erroneous reasons of BERT-base and BERT-ADA on the MAMS test dataset, and the reasons that lead to a better performance of BERT-ADA. Input reduction is used to extract important words in erroneous samples. Sampled incorrectly predicted sentences are shown in Figure 3. We will illustrate them by the reference number.

RD1, RD2: Lack restaurant-domain knowledge. In both original sentences, BERT-ADA gives right answers but BERT-base fails. In restaurant review context, "too sweet", "over 45 mins" are negated sentiments. As BERT-ADA is trained on restaurant-domain corpus, it contains those knowledge that BERT-base lacks.

GR1, GR2: Lack of general review-domain knowledge. "not...again", "below average" are often found in reviews context, which means negation. Hence, BERT-ADA gives right answers again.

MA1: Affected by muti-aspect muti-sentiment sentence. "superior" describes aspect "attitude" rather than "food". BERT-base is confused by this fact and that is why MAMS dataset is challenging.

AS1, AS2, AS3: Ambiguous sentiment. Generally, "addictive" shows positive sentiment, but "greasy" shows negated sentiment. BERT-base and BERT-ADA models assign different weights to these words, which leads to different classifications. Sentiment of "price was unbelievable" is ambiguous, we may need more contexts.

SR1: Lack of syntax rules. "not only...but..." implies the two parts should have similar sentiments. Even "unbelievable" is ambiguous, "COOL" and "spectacular" are positive, which helps BERT-ADA

Models		SemEval-14		MAMS	
		Accuracy	F1-score	Accuracy	F1-score
GloVe	LSTM	0.7268	0.5301	0.5122	0.3712
	ATAE-LSTM	0.7491	0.6033	0.7028	0.5259
	CapsNet	0.8079	-	0.7978	-
	CapsNet-BERT	0.8593	-	0.8339	-
BERT	BERT-base	0.8492	0.7693	0.8406	0.8356
	BERT-ADA	0.8714	0.8005	0.8473	0.8419

Table 2: Results on SemEval-14 and MAMS Dataset.

give a correct classification. Note this is difficult because even Gold test set makes a mistake.

Sentence and **Aspect**	Ref.	Reduced Words	Gold	Base	ADA
The mango salsa with fish cake was too sour, the **apple suace** for pork chop too sweet.	/	/	↓	↑	↓
	RD1	too sweet	/	—	—
We had reservations and when we showed up the manager told us the **wait** was over 45 mins.	/	/	↓	—	↓
	RD2	over	/	—	—
As for the food, brunch was average, I would not get the same **dish** again, and they were slow to serve us.	/	/	↓	—	↓
	GR1	not	/	—	—
I find the attitude of the managers to be appallingly superior, and the **food** below average, at sky-high prices.	/	/	↓	—	↓
	GR2	below	/	—	—
	MA1	superior	/	↓	↓
What do you like more, completely addictive patties of greasy **beef**, or glasses of cheap refreshing beer to wash it down?	/	/	↑	↑	↓
	AS1	addictive	/	↓	↓
	AS2	greasy	/	↑	↑
I will be back not only because the **price** was so unbelievable but the atmosphere was just plain COOL and the food was spectacular.	/	/	↓	↓	↑
	AS3	unbelievable	/	—	↑
	SR1	not only...but	/	↑	↑

Figure 3: Shown are samples from MAMS test data where the BERT-base or BERT-ADA predicts polarity incorrectly. The reduced words are keywords that affects the prediction outcome if removed. Notes: 1) words with bold and underline mean aspects; 2) "↑" means positive, "-" means neutral, "↓" means negative; 3) signal in red means wrong polarity; 4) words in purple represent reduced words

5 Conclusion

In our project, we have conducted aspect-based sentiment analysis. Specifically, we performed experiments on the task of aspect-term sentiment classification. We implemented LSTM and ATAE-LSTM as baseline, as well as the vanilla BERT-base model and BERT-ADA that is pretrained on Yelp restaurant reviews. These models are applied to a more challenging MAMS dataset. For MAMS, each sentence contains multiple aspects with different sentiment polarities, which is more practical when used in real business settings.

Two BERT-based models generalize much better than LSTM models, and the state-of-the-art BERT-ADA model on the SemEval dataset performs best on MAMS, beating the CapsNet-BERT model. We further gave a robust error analysis to find out the reasons for those erroneous classifications, and why BERT-ADA outperforms other models.

For future work, we plan to explore the possibilities of applying other neural methods like CNN on top of the domain-trained BERT-ADA to improve the prediction accuracy. Another interesting and possible direction would be to combine aspect-term extraction and sentiment classification, and build a unified model in an end-to-end fashion.

References

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

Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, page 231–240, New York, NY, USA. Association for Computing Machinery.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. volume 2, pages 49–54.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. arXiv:1804.07781.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 504–515, Florence, Italy. Association for Computational Linguistics.

Binxuan Huang and Kathleen M. Carley. 2018. Parameterized convolutional neural networks for aspect level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1091–1096. Association for Computational Linguistics.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent Twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 151–160, Portland, Oregon, USA. Association for Computational Linguistics.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 6280–6285. Association for Computational Linguistics.

Tony Mullen and Nigel Collier. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, volume 4, pages 412–418. Association for Computational Linguistics.

Thien Hai Nguyen and Kiyoaki Shirai. 2015. PhraseRNN: Phrase recursive neural network for aspect-based sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2509–2514, Lisbon, Portugal. Association for Computational Linguistics.

Maria Pontiki, Dimitrios Galanis, John Pavlopoulos, Haris 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. Association for Computational Linguistics.

Delip Rao and Deepak Ravichandran. 2009. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, page 675–682, USA. Association for Computational Linguistics.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2019. Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification. arXiv:1908.11860.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *neural information processing systems*, pages 5998–6008.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 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, Austin, Texas. Association for Computational Linguistics.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. arXiv:1904.02232.