# Multi-Aspect Attention Model for Aspect-based Sentiment Classification Using Deep Learning

Win Lei Kay Khine, Nyein Thwet Thwet Aung
*Faculty of Information Science, University of Information Technology, Yangon, Myanmar*
*winleikkhine@uit.edu.mm, nyeinthwet@uit.edu.mm*

## Abstract

*Aspect-based sentiment classification (ABSC) (also called fine-grained sentiment analysis) is an essential and challenging task among sentiment analysis tasks. Aspect level sentiment analysis overcomes the limitation of the document and sentence level when multiple aspects appear in a review. Recently, neural network approaches like LSTM has achieved better results in ABSC than traditional machine learning algorithms. Aspect extraction and polarity detection for specific aspects in the review becoming an important task in aspect level sentiment analysis. Therefore, the paper aims to predict the sentiment polarity of each aspect in the review into 3 classes by developing the multi-aspect attention (MAA) model and combine it with the BiLSTM neural network, called MAA-BLSTM. Unlike unidirectional LSTM, the BiLSTM network runs the input in two directions: forward and backward, therefore, it can understand the context better than LSTM. In this paper, the hyperparameter tuning approach is also considered to be the high-performing model. Experiments are tested on restaurant and laptop data from SemEval (2014, 2015, and 2016) datasets and compare the result with other LSTM-based methods. Finally, the result proves that the accuracy of the proposed sentiment model reaches 89.9 on the restaurant dataset and 82.1 on the laptop dataset which are higher than other LSTM approaches.*

**Key Words-** Aspect-based sentiment classification (ABSC), Deep Learning, Neural Networks, Multi-Aspect Attention (MAA) Model, BiLSTM, Hyperparameter tuning

## 1. Introduction

Many researchers have analyzed the sentiment classification at three levels: document level, sentence level, and aspect or entity level. Among them, aspect level sentiment analysis becomes an essential one in the sentiment classification task because it considers different polarities of each aspect mentioned in the review. Therefore, this paper proposes to analyze the sentiment analysis at the aspect level. The goal of ABSC is: (1) to extract the target aspects from the given review and (2) to predict the sentiment polarity of each aspect.

For example, "*The price is reasonable although the service is poor*". The sentence talks about positive polarity for aspect price and negative polarity for aspect service in one sentence. Sometimes, the customers can write a review of products or services on social media how they like or dislike it. If there are many aspects in review, how to correctly predict the sentiment polarity becomes a challenge in sentiment analysis.

Many researchers typically use machine learning approaches for sentiment analysis in a supervised manner. Nowadays, deep learning algorithms like recurrent neural networks (RNN) are popular for sequence models and it gains state-of-the-art results in sentiment analysis tasks. To overcome this issue, the BiLSTM network is used because it can avoid the vanishing gradient or exploding problem and can handle learning long-term dependency [5].

The main aim of this paper is to develop the combination of the BiLSTM network with the multi-aspect attention (MAA) model. BiLSTM is an extension of LSTM and GRU recurrent neural network and it is used for a sequential learning problem. It manages the input information in two ways: one from past to future and the other from future to past (called forward and backward) in a specific time frame.

The attention mechanism is a recent trend in deep learning and is widely used in image processing for paying visual attention to different regions of an image. It is also widely applied in machine translation by paying attention to correlate words in one sentence. In this paper, the attention mechanism is used to solve the problem of the ABSC task. It captures the most important contexts from the input sentence and gives attention to the aspect of a specific target by calculating the attention weight. There are four main objectives in this paper: (1) extract the aspects and predict the relevant polarities in the review (2) use hyperparameter tuning approach with BiLSTM network to prevent from the vanishing gradient problem that is faced in training the neural network model, (3) develop multi-aspect attention (MAA) model for capturing the most relevant aspects from the given review, and (4) compare the results with other LSTM-based methods.

The remainder of the paper is organized as follows: Section 2 presents the related work of aspect level sentiment analysis by using LSTM-based methods. Section

3 discusses the methodology: word embedding layer, BiLSTM layer, attention mechanism, multi-aspect attention (MAA), and sentiment classification. Section 4 emphasizes experiments such as about dataset, hyperparameter tuning, results, evaluation, and environmental setup. Finally, the paper concludes with Section 5.

## 2. Related Work

This section discusses the research works that are related to aspect level sentiment analysis. Many research works are proposed by using machine learning classifiers like Naïve Bayes, SVM, and Maximum Entropy. These classifiers are built on manually created features. In this paper, we propose the ABSC task with the neural network-based method which is end-to-end training without feature engineering and it can produce state-of-the-art performance.

In [1], the authors present an attention-based LSTM Network for aspect-level sentiment classification. They use an aspect term embedding to generate an attention vector to concentrate on different parts of the sentences and compute the attention weights. The results show the accuracy of their approach is higher than LSTM and T-LSTM methods on three-way classification and binary classification. They tested the experiments on restaurant data of SemEval 2014 and achieves 77.2% for 3way classification and 90.0% for binary classification.

Tang et al. [2] consider the target information and achieved good performance in target-dependent sentiment classification. However, Target-Dependent LSTM (TD-LSTM) is not good enough and cannot capture the interactions between target and context words. Based on its consideration, they develop a target-connection LSTM (TC-LSTM) which obtained a target vector by averaging the vectors of words the target phrase contains. The experiments are conducted on the Twitter dataset and compare the results with the adaptive recursive neural network, lexicon-enhance neural network, and feature-based SVM.

The authors in [3] present the deep learning approach for the multilingual aspect-based sentiment analysis in SemEval 2016. They use an extension of CNN by Collobert et al. (2011) and especially focus on aspect category extraction and sentiment polarity. They concatenate an aspect vector into word embedding and show the result with accuracy in a multilingual setting.

In [10], the authors experimented on the restaurant data of SemEval 2014, 2015, and 2016, and get 88.91%, 65.97%, and 76.42% on the F1 score. They use Multi-task learning by combining CNN and BiLSTM neural networks.

The authors in [11] presented the ABSC task by using BiGRU, aspect attention, and sentiment attention on restaurant data. They achieve an accuracy of 85.1% for 3- way classification and 91.3 for binary classification.

## 3. Methodology

In this section, the methodology for aspect sentiment classification is presented. It contains four main components: (1) word embedding layer, (2) the BiLSTM layer is to generate the sentence representation, (3) multi-aspect attention (MAA) model is to pay attention over context words, and (4) sentiment classification will be discussed. The overall system architecture is shown in Figure 1.

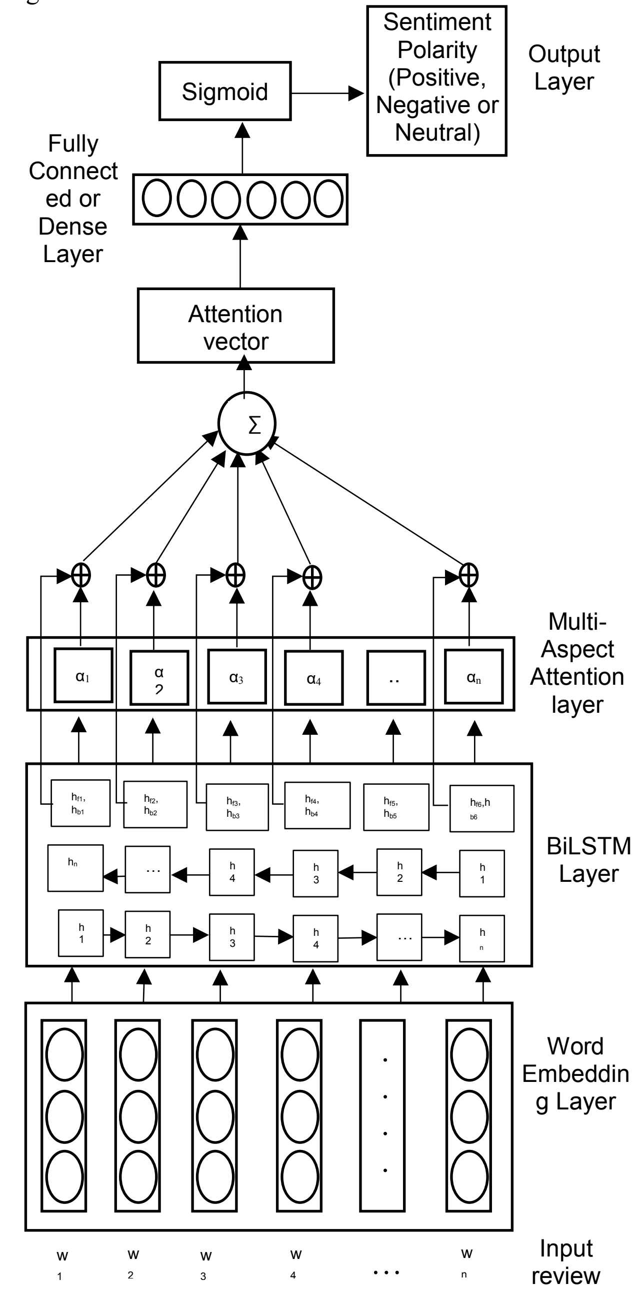


**Figure 1. The system architecture of the MAA-BLSTM model for the ABSC task.**

### 3.1. Word Embedding Layer

In the word embedding layer, the paper uses 100 and 300-dimension pre-trained GloVe word embedding matrix [6]. Let, $\mathbb{L} \in \mathbb{R}^{d_{emb} \times |V|}$, where |V| is the vocabulary size and $d_{emb}$ is the embedding dimension of word vectors. Each word is mapped to its corresponding embedding vector space through this layer. The output of word embedding is word vectors for the review sentences.

### 3.2. BiLSTM Layer

After creating the embedding layer, the word vectors are fed into the BiLSTM layer to compute the hidden states of input embedding. Unlike unidirectional LSTM, BiLSTM uses two networks: one from past to future and the other from future to past. The advantage is that it can avoid the gradient vanishing or exploding problem and solves the long-term dependency problem.

In the forward LSTM, the input is feed as a sequence of vectors, for example, $s = [v_1; v_2; \ldots; v_n]$ and it generates the hidden state as output,

$$\overrightarrow{h_s} = \overrightarrow{LSTM}\ ([v1;\ v2;\ \ldots;\ vn]) \qquad (1)$$

The backward LSTM also calculates the reverse hidden state.

$$\overleftarrow{h_s} = \overleftarrow{LSTM}\ ([v1;\ v2;\ \ldots;\ vn]) \qquad (2)$$

Final output representation is the concatenation of forward and backward LSTM and can be represented as**:**

$$h_s = \overrightarrow{h_s} \oplus \overleftarrow{h_s} \qquad (3)$$

In the BiLSTM layer, we use 128 neuron units as the optimal. BiLSTM layer with 128 hidden neurons produces 128 features for the next layer. Finally, the two output vectors (forward and backward) are concatenated and sent to the MAA layer.

### 3.3. Attention Mechanism

In recent years, attention is a popular and useful tool in deep learning. It was born for memorizing long source of sentences in neural machine translation (NMT). Therefore, the paper applies the attention mechanism for sentiment classification.

Attention works as encoder-decoder architecture and it addresses the limitations of encoder-decoder problems on long sequences. The encoder processes the input sentences and compresses the information into a context vector of fixed-length and the decoder initializes with the context vector to transmit the transformed output. The attention weight is produced between the context vector and the input sequences. The alignment between the source and the target is controlled by the context vector. The context vector consumes 3 parts of information (1) encoder hidden states (2) decoder hidden states and (3) alignment between source and target.

### 3.4. Multi-Aspect Attention (MAA)

The input to the MAA is the hidden representation of the previous layer, BiLSTM. The BiLSTM encoder works as the concatenation of two encoder states (forward hidden state $\overrightarrow{h_s}$ and backward hidden state$\overleftarrow{h_s}$ ). The BiLSTM layer generates the encoded representation of the sentences and the aspect targets. After then, these representations are fed to the MAA layer, In the MAA layer, the attention weights for the given review are calculated. The output representation of the sequence encoding is attention weight such as α1, α2, α3, and so on. Computing attention weight is a weighted sum of word representations obtained from the BiLSTM encoder layer based on their attention weights. The attention weight $\alpha_{js}$ for each context word is shown in equation (4).

(Attention weight)

$$\alpha_{js} = \frac{\exp(score(hj,hs))}{\sum_{k=1}^{N}\exp(score(hk,hs))} \qquad (4)$$

It captures the interactions between aspects and context words. Using the attention in the ABSC task enables easier learning with high quality because aspect information plays a vital role in aspect level sentiment classification. It does not focus on everything and focus only on specific aspects and polarity. Examples of multi-aspects and their relevant polarities from the datasets are shown in Table 1.

Firstly, the model calculates the attention scores from the hidden state of representation $\{h_1, h_2, h_3, \ldots, h_i, \ldots, h_N\}$. After that, the attention scores are passed into the Softmax activation function to produce the attention weights such as α1, α2, α3, and so on. Finally, the attention weights and hidden states are to produce the context vector, $c_t$.

(Context vector)

$$c_t = \sum_{j=1}^{N}\alpha_{js}h_j \qquad (5)$$

The context vector, $c_t$ is a sum of the hidden state of the input sequence, weighted by alignment scores. Finally, the attention vector is calculated in which tanh is a non-linear activation function, W is the weight matrix and b is the bias.

(Attention vector)

$$a_j = \tanh\ (\mathrm{W}[c_t;\mathrm{hj}\ ]+\mathrm{b}) \qquad (6)$$

**Table 1. Example sentences of restaurant and laptop data and its aspects from the datasets.**

| Domain | Sentences | Aspects | Polarity |
|---|---|---|---|
| Restaurant | We went here for lunch a couple of weeks ago on a Saturday, and I was thoroughly impressed with the food. | lunch<br>food | neutral<br>positive |
| Laptop | One night I turned the freaking thing off after using it, the next day I turn it on, no GUI, screen all dark, power light steady, hard drive light steady and not flashing as it usually does. | GUI<br>screen<br>power light<br>hard drive light | negative<br>negative<br>neutral<br>negative |

### 3.5. Sentiment Classification

After the MAA layer, the output representation is forwarded to the dense or fully-connected layer via activation function. This layer maps the output of the previous layer to the desired output size. This layer learns to keep the relevant information for sentiment prediction for the target aspects and forget irrelevant data. If the probability of polarity for an aspect surpasses a threshold, the aspect will be assigned to the respective sentiment class. The final task of ABSC is to classify the sentiment polarity, the model accepts $a_y$ as input features. The final predicted sentiment polarity of an aspect target is the label with the highest probability. For the classification, we use a sigmoid activation function that can output the values in a range between 0 and 1.

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \,.\, log(p(y_i)) + (1 - yi)\,.\ \ log(1 - p(y_i)) \qquad (7)$$

We trained our model to minimize the cross-entropy loss with L2 regularization.

Table 2 and Table 3 shows some examples of aspect category and its target aspect used in the paper.

## 4. Experiments

This section discusses about the dataset, hyperparameter tuning, experimental result, and environment setup.

### 4.1 Dataset

The ABSC model uses restaurants, and laptop data from SemEval 2014 Task 4 [7], 2015 Task 12 [8], and 2016 Task 5 [9]. All the data from these datasets are in .xml format and we merge these all datasets into one for training and testing the model. The detailed datasets used in this paper are described in Tables 4, 5, and 6.

**Table 2. Samples of Targeted Aspects and Aspect Categories in restaurant datasets.**

| Aspect Category | Aspect Term |
|---|---|
| AMBIENCE#GENERAL | Atmosphere, interior décor, style |
| FOOD#PRICES | Burger, salad, sandwich |
| FOOD#STYLE_OPTIONS | Desserts, presentation of food |
| RESTAURANT#GENERAL | Address, bar, café |
| SERVICE#GENERAL | Host, management, service, staff |

**Table 3. Sample Aspect Categories and Aspects Terms in laptop datasets.**

| Aspect Category | Aspect Term |
|---|---|
| BATTERY#QUALITY | Battery life, battery |
| DISPLAY#GENERAL | Appearance, resolution |
| LAPTOP#DESIGN_FEATURES | Color, style |
| LAPTOP#GENERAL | Brand, quality, startup, windows |
| LAPTOP#PRICES | Price, cost, price tag |
| OS#GENERAL | System, operating system, BIOS update, |
| WARRANTY#GENERAL | Service, warranty, warranty service |

**Table 4. Restaurant and laptop data from SemEval 2014 Task 4**

| Dataset | Positive | Neutral | Negative | Total |
|---|---|---|---|---|
| Training (Restaurant) | 2164 | 637 | 807 | 3608 |
| Testing (Restaurant) | 728 | 196 | 196 | 1120 |
| Training (Laptop) | 994 | 464 | 870 | 2328 |
| Testing (Laptop) | 341 | 169 | 128 | 638 |

**Table 5. Training and testing data of restaurant and laptop from SemEval 2015 Task 12.**

| Domain | Training | Testing | Total |
|---|---|---|---|
| Restaurant | 1654 | 845 | 2499 |
| Laptop | 1974 | 949 | 2923 |

**Table 6. Training and testing data of restaurant and laptop data from SemEval 2016 Task 5.**

| Domain | Training | Testing | Total |
|---|---|---|---|
| Restaurant | 2000 | 676 | 2676 |
| Laptop | 2500 | 803 | 3303 |

**Table 7. Hyperparameter tuning on the proposed system on the restaurant dataset with different lengths of the input sentence.**

| Pretrained Word embedding | Batch _size | Epochs | Activation | Optimizer | Loss | Train_acc | Val_Acc | Test_acc | Remark |
|---|---|---|---|---|---|---|---|---|---|
| Glove 6B 100 dimension | 32 | 8 | Sigmoid | Adam | BCE | 95.2 | 85.2 | 85.3 | Dropout=0 |
| Glove 6B 100 dimension | 16 | 8 | Sigmoid | Adam | BCE | 96.9 | 85.8 | 85.1 | Dropout=0 |
| Glove 6B 100 dimension | 32 | 15 | Sigmoid | Adam | BCE | 98.0 | 84.2 | 84.6 | Dropout=0 |
| Glove 6B 100 dimension | 32 | 15 | Sigmoid | Adam | BCE | 98.0 | 84.9 | 84.5 | Dropout=0.2 |
| Glove 6B 100 dimension | 32 | 15 | Sigmoid | SGD | BCE | 83.4 | 82.9 | 81.9 | Changing Optimizer |
| Glove 42B 300 dimension | 64 | 10 | Sigmoid | Adam | BCE | 97.9 | 90.4 | **89.9** | Changing dimension to 300 and Dropout=0.2 |
| Glove 42B 300 dimension | 64 | 15 | Sigmoid | Adam | BCE | 99.2 | 90.0 | 89.8 | Dropout=0.2 |
| Glove 42B 300 dimension | 64 | 15 | Sigmoid | Adam | BCE | 98.4 | 89.3 | 89.4 | Dropout=0.5 |
| Glove 42B 300 dimension | 64 | 15 | Sigmoid | Adam | BCE | 96.7 | 89.1 | 89.3 | Dropout=0.8 |
| Glove 42B 300 dimension | 64 | 15 | Sigmoid | Adam | BCE | 98.7 | 89.9 | 89.8 | Add Flatten Layer, and Dropout=0.5 |
| Glove 42B 300 dimension | 64 | 15 | Sigmoid | Adam | BCE | 98.6 | 89.2 | 89.2 | Add Flatten Layer, and Dropout=0.2 |

### 4.2 Hyperparameter Tuning

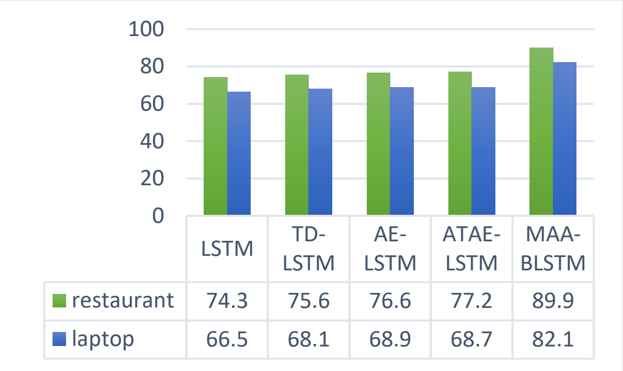
Hyperparameter tuning is an essential part of training the neural network models and that governs the entire training process. It is applied in all layers to produce the effective neural sentiment model. The model tunes the different hyperparameters such as pre-trained word embedding with 100 and 300 dimensions, batch size, number of epochs, dropout rate, optimizer, loss, and activation function on the BiLSTM network. Performance evaluation for hyperparameter tuning of the restaurant dataset is shown in Table 7. Like the restaurant data, laptop data is also tested with the same hyperparameters. The optimal hyperparameters for both restaurant and laptop data for MAA-BLSTM are shown in Table 8.

### 4.3 Result and Evaluation

The model shows the performance with the accuracy measurement. Firstly, it trains and tests the model with different lengths of words using the BiLSTM network on different dimensional embedding matrix like 100 and 300, different epochs and batch size, optimizers, and activation functions. When we train the model, we face an overfitting problem. To overcome this problem, we use the different dropout rate range from 0 to 1 and choose the optimal dropout rate as 0.2. The comparative results of the overall accuracies of baseline models and our model for restaurant and laptop data are shown in Figure 2. Our MAA-BLSM model gets 89.9% accuracy in restaurant data and 82.1% in laptop data which are the highest among other LSTM-based models. For the experimental setup, the Keras framework is used and Tensorflow is used as the backend. The model is implemented using Python. Experiments are tested on Google Colaboratory that allows us to write and execute Python with no configuration required and free access to GPUs and TPUs on Google's cloud server.

**Table 8. Optimal Hyperparameter for both restaurant and laptop domain after training and testing the model.**

| | |
|---|---|
| Pretrained Embedding | Glove (300 dimensions) |
| Max Size | 300 |
| Train/Dev/Test | Training 90% (validation 10% include) and Testing 10% |
| Batch Size | 64 |
| Epochs | 10 |
| Activation | Sigmoid |
| Optimizer | Adam |
| Learning rate | 0.001 |
| Momentum | 0.1 |
| Dropout rate | 0.2 |
| Loss function | BCE (Binary CrossEntropy) |



| | LSTM | TD-LSTM | AE-LSTM | ATAE-LSTM | MAA-BLSTM |
|---|---|---|---|---|---|
| 🟩 restaurant | 74.3 | 75.6 | 76.6 | 77.2 | 89.9 |
| 🟦 laptop | 66.5 | 68.1 | 68.9 | 68.7 | 82.1 |

**Figure 2. Comparison of Accuracies on Multi-aspect attention BiLSTM (MAA-BLSTM) model with other LSTM-based methods by using the same hyperparameter tuning on the SemEval's restaurant and laptop datasets.**

## 5. Conclusion

In conclusion, this paper develops a sentiment model for aspect-based sentiment classification by combining the multi-aspect attention (MAA) model and BiLSTM. The result shows that adding MAA to BiLSTM improves accuracy than other models. Hyperparameters tuning is applied to control the behavior of the training process and to get high-performance results. This paper implemented on the restaurant, and laptop data from SemEval datasets. Finally, experiments show that the MAA-BLSTM with a well-trained hyperparameter tuning approach gets higher accuracy than baselines LSTM models.

## 6. References

[1] Y. Wang M. Huang, L. Zhao, and X. Zhu, "Attention-based LSTM for Aspect-level Sentiment Classification", *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Austin, Texas, November 2016, pp. 606-615.

[2] D. Tang B. Qin, X. Feng, T. Liu, "Effective LSTMs for Target-Dependent Sentiment Classification", Proceeding of COLING 2016, the 26th International Conference on Computational Linguistics, Osaka, Japan, December 11-17, 2016, pp. 3298-3307.

[3] S. Ruder,P. Ghaffari, J.G. Breslin, "INSIGHT-1 at SemEval-2016 Task 5: Deep Learning for Multilingual Aspect-based Sentiment Analysis", Proceedings of SemEval-2016, ©2016 Association for Computational Linguistics, San Diego, Califonia, June 16-17, 2016, pp. 330-336.

[4] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, "Natural Language Processing (almost) from Scratch", Journal of Machine Learning Research, August 12, 2011, pp. 2493-2537.

[5] S. Hochreiter, J. Schmidhuber, "LONG SHORT-TERM MEMORY". *Neural Computation* Vol. 9, No. 8, 1735-1780 (1997).

[6] J. Pennington, R. Socher, C. D. Manning, "Glove: Global vectors for word representation". *Proceedings of the 2014 conference on empirical methods in natural language processing* (EMNLP). Association for Computational Linguistics, Doha, Qatar, October 2014, pp. 1532–1543.

[7] http://alt.qcri.org/semeval2014/task4/

[8] http://alt.qcri.org/semeval2015/task12/

[9]http://alt.qcri.org/semeval2016/task5/index.php?id=data-and-tools

[10] W. Xue, W, Zhou, T. Li, Q. Wang, "MTNA: A Neural Multi-task Model for Aspect Category Classification and Aspect Term Extraction On Restaurant Reviews", Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2), pp. 151-156.

[11] J. Cheng, S. Zhao, J. Zhang, I. King, X. Zhang, H. Wang, "Aspect-level Sentiment Classification with HEAT (HiErarchical ATtention) Network", CIKM '17: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, November 2017, pp. 97-106.s