Aspect-Based Sentiment Analysis of Specific Targets

Lei Hong *

School of Mathematical Sciences, Anhui University, Hefei, Anhui province, 230601, China

* Corresponding Author Email: a01914064@stu.ahu.edu.cn

Abstract. Different from traditional sentiment analysis tasks, Aspect-Based Sentiment Analysis (ABSA) aims to automatically identify the emotional polarity of different aspects in the same sentence, which helps to mine users' more delicate emotional expressions for different targets and has become a research hotspot in the field of natural language processing in recent years. Thanks to the rapid development of attention-based deep neural network models, the accuracy of aspect sentiment analysis has continuously made breakthroughs. However, existing works pay little attention to the performance bounds of different application scenarios, such as different text topics or text lengths. In response to the above issues, this paper conducts a comparative analysis of several classic end-to-end neural network models to explore topic-level sentiment polarity analysis. By quantitatively evaluating the accuracy and F_1 score of each model at different fixed text lengths, this paper summarizes the strengths, weaknesses, and different aspects of each model's efficiency in use. In addition, this paper also discusses the existing problems with these models and puts forward suggestions for improvement, which can provide some new insights into the research of ABSA.

Keywords: Aspect-based sentiment analysis; deep learning.

1. Introduction

In this era of digitalization, the internet is intricately linked with the daily lives of individuals. Through channels like social media, product reviews, and blog posts, people express their viewpoints and emotions. These data converge into vast volumes of information, encompassing sentiments and attitudes towards various topics such as products, services, events, policies, etc. Understanding and analyzing the emotional content in these texts can be of great help for product optimization and decision-making support [1]. The growth of user-generated content on the internet provides a sufficient data base for sentiment analysis, and also promotes sentiment analysis as an important tool for extracting emotional state information. In this context, sentiment analysis has gradually become a popular research task in the field of natural language processing, attracting a lot of research interest.

According to text granularity, sentiment analysis can be divided into coarse-grained and fine-grained hierarchical analyses. Previous efforts focused on the coarse-grained task, which mainly comprises chapter and sentence-level sentiment analyses [2]. They analyze the sentiment of given texts, differing only in text length [1]. However, due to distributional differences in individual expressive habits, positive and negative emotional polarities are often mixed when a single text involves multiple topics at the same time, which results in great difficulty in ascertaining the real emotions of the target text if topic information is disregarded [3]. For example, in a user's restaurant review, the text covers both the food and service, each with opposite emotional tones [4]. When conducting sentiment analysis in the text, it becomes challenging to determine whether the overall sentiment is positive, negative, or even neutral. As a result, researchers initiated more detailed sentiment analysis, ultimately establishing the concept of aspect-level tasks in 2010 [5]. Compared to common sentiment analysis, aspect-level evaluation constitutes a precise and challenging task within sentiment classification, which aims at predicting the emotional valence of specific targets in a given text [6]. In practical scenarios, individuals need to identify not only the points of view in a document or sentence, but also the objects towards which emotions or evaluations are directed, and the specific feelings expressed towards these objects [7-8]. ABSA, as a deeper sentiment analysis, can mine users' more delicate emotional expressions for different targets and has become one of the research hotspots in the field of natural language processing in recent years.

In some past studies, traditional machine learning methods have achieved a lot of success in sentiment analysis tasks associated with targets. However, these methods usually require a lot of preprocessing and complex feature engineering, as well as dependency analysis on the input text. In recent years, deep learning has made major breakthroughs in text classification tasks, and attention-based deep neural network models have become mainstream ABSA frameworks [9]. Although existing research has greatly improved the performance of aspect sentiment analysis, little attention has been paid to the performance ceilings of multiple methods for various application situations. Significant differences in the results of a method can be easily observed when analyzing text reviews of different topics or lengths. Exploring the application boundaries of different methods can provide a decision-making basis for method selection in practical applications, which has important application value.

To alleviate the above issues, this paper explores the performance of representative aspect sentiment analysis methods in detail. Specifically, this paper first reviews the basic theories of different aspects of sentiment analysis methods, including their design ideas, key steps, network structure, etc. Second, this paper quantitatively compares the outcomes obtained from different methods on a shared resource and summarizes the advantages and disadvantages of different methods. Finally, the paper also discusses the existing adjustments in aspect sentiment analysis and looks forward to the future development direction of this field.

2. Method

2.1. Related work for ABSA

In the context of sentiment classification at the topic level, previous studies can be broadly classified into two categories. One category comprises cascading techniques that rely on features extracted manually and algorithms from machine learning. In the research conducted by Nasukawa et al. [10], the first step is to carry out dependency syntactic analysis of sentences, followed by applying predetermined rules to figure out the psychological mood for each subject matter. Jiang et al. [11] suggested establishing several topic-dependent features on the foundation of sentence grammar structures. These features, along with other textual content features, are employed as input to an SVM classifier, thereby enhancing its precision.

Another class of endeavor utilizes end-to-end neural network models. The use of neural network models enables automated feature learning, eliminating the need for substantial manual engineering or rule crafting. Within these models, most are constructed using LSTM network architectures. LSTM is skilled at managing sequential data, such as text and capturing enduring dependencies. Tang et al. [12] introduced the TD-LSTM approach, which aims to model the topic context by employing two separate LSTM networks. The outputs of these two LSTM networks are then concatenated to predict sentiment. Additionally, in recent years, extensive use of attention mechanisms in conjunction with LSTM has effectively enhanced network performance. Wang et al. [13] proposed two classic foundation frames. In AE-LSTM, the texts undergo an initial LSTM layer of networks processing to derive hidden unit outputs for every word. Subsequently, a group of attention values is calculated by merging with topic vector representations, assigning weightages to different portions of the sentence in order to arrive at the conclusive vector representation for sentiment classification. By contrast, ATAE-LSTM prioritizes the impact of topic vector representation. The design constructs the text representation by combining the vector representation of the topic with the expression results of each word.

2.2. Model presentation

In this section, we first introduce the most representative methods for analyzing sentiment at the aspect level in detail, including: the Interactive Attention Networks (IAN) model [14], the Attention Over Attention (AOA) model [4], and the Attention Encoder Networks (AEN) model [15].

2.2.1 Interactive Attention Networks

The architecture of IAN is displayed in Figure 1 and is comprised of four segments [14].

(1) Word Embedding Module: After disambiguation, a sentence can be denoted as $s = [w_1, w_2, \cdots, w_n]$. For all topics, it can be denoted as $t = [w_1, w_2, \cdots, w_m]$, and for this task, there are ten topic words, so m = 10. Using pre-trained word vectors, each word w_i is mapped to a low-dimensional vector $v_i \in R^{d_w}$ with real values. Sentences and topics can be passed through a word-embedding layer to obtain two vector groups $s = [v_1; v_2; \cdots; v_n] \in R^{n \times d_w}$ and $t = [v_1; v_2; \cdots; v_m] \in R^{m \times d_w}$. The word embedding layer can be used to embed words and topics into the module.

(2) Bidirectional LSTM Module: Word vectors s and t are processed separately through a bidirectional LSTM layer for the acquisition of latent representations pertaining to themes and lines. The output is the mixing of hidden unit outputs from the bidirectional LSTM, resulting in hidden states $h_s \in R^{n \times 2d_h}$ and $h_t \in R^{m \times 2d_h}$.

(3) Interactive Attention Module: For the obtained h_s and h_t from the previous step, average pooling is applied, resulting in $s_{avg} = \sum_{i=1}^{n} \frac{h_s^i}{n}$ and $t_{avg} = \sum_{i=1}^{m} \frac{h_t^i}{m}$. A set of attention weights is computed from t_{avg} and the sentence's hidden states h_s, where each weight is defined as $\alpha_i = \frac{\exp(\gamma(h_s^i, t_{avg}))}{\sum_{j=1}^{n} \exp(\gamma(h_s^j, t_{avg}))}$. Different portions of the phrase are weighted using the attention values, resulting in the final vector representation of the sentence $s_r = \sum_{i=1}^{n} \alpha_i h_s^i$. Simultaneously, another set of attention weights is calculated using s_{avg} and the hidden states h_t of topics. Each weight, denoted as $\beta_i = \frac{\exp(\gamma(h_t^i, s_{avg}))}{\sum_{j=1}^{m} \exp(\gamma(h_t^j, s_{avg}))}$, is computed from this set of attention scores to evaluate various topics, resulting in the vector display of the topics $t_r = \sum_{i=1}^{m} \beta_i h_t^i$. The computation formula for both sets of weights is defined as $\gamma(i, j) = \tanh(i \cdot W_a \cdot j^T + b_a)$.

(4) Classification Module: Concatenate the final vector representations t_r and s_r of the topic and sentence received in the previous stages, and feed them into a fully linked two-layer network that generates a rating outcome. Four-dimensional outcome findings are readily accessible for every target, representing the probabilities of the sentence not containing that topic, and the probabilities of the sentence having a positive, neutral, or negative sentiment polarity within that topic. Apply the softmax function to the output results corresponding to each topic in order to ascertain whether the sentence lacks representation of that topic or to measure its polarity inside that target-oriented surroundings.

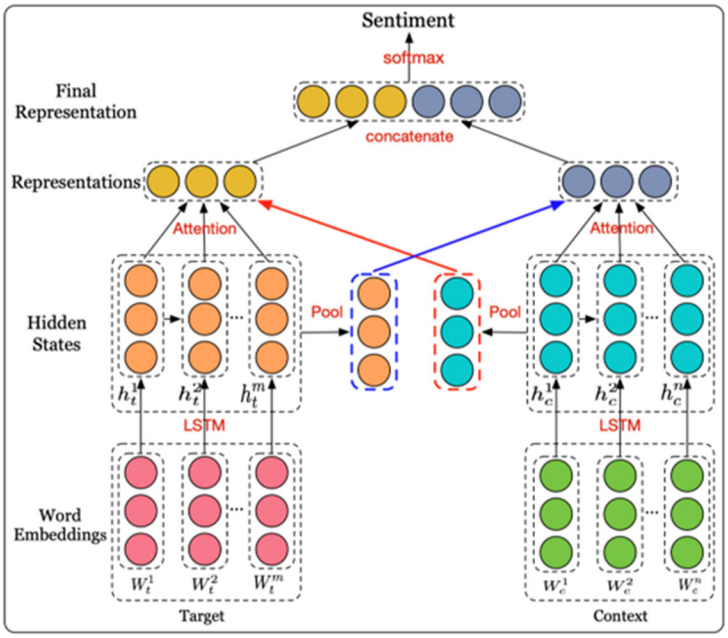

Fig. 1 The comprehensive structure of the IAN model.

2.2.2 Attention Over Attention

The structure of the AOA model [4] is illustrated in Figure 2 and is divided into four components. The first two modules resemble the IAN model, with the innovation lying in the attention module. The interaction matrix $I = h_s \cdot h_t^T$ is first calculated for the h_s and h_t obtained in the previous step, where each element I_{ij} within the matrix signifies the association between a particular word w_i in the sentence and a particular topic w_j. Softmax operations are applied to each column of matrix I to acquire the attention weightings α for aspect-to-sentence, where $\alpha_{ij} = \frac{\exp(I_{ij})}{\sum_i \exp(I_{ij})}$. The value of each α represents which words are relatively more significant for a specific topic. Ultimately, the sentence vectors representing the perception of various topics are calculated based on the weights α and the hidden semantic representation h_s of the sentence, denoted as $x = \alpha^T \cdot h_s \in R^{m \times 2d_h}$. The acquired sentence representation vector x is flattened and then fed into a fully connected network similar to the previous one to produce classification results. Finally, softmax is applied to the obtained outcomes for each subject and then output.

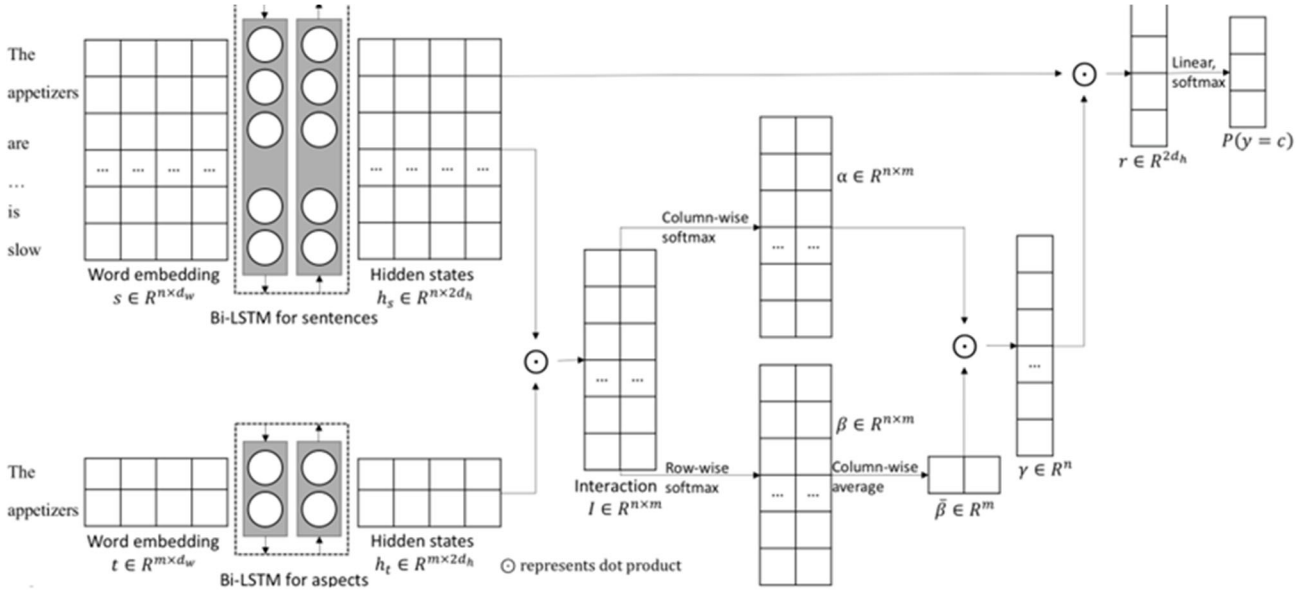

Fig. 2 The comprehensive structure of the AOA model.

2.2.3 Attention Encoder Networks

The architecture of the third model, the AEN model [15], is depicted in Figure 3. Within this model, the first three modules of the framework are the same as those in AOA. In total, three output vectors are obtained as inputs for the subsequent module: the concealed status h_s attributed to the sentence, h_t of the topic, and the topical condition of interaction $x = \alpha^T \cdot h_s \in R^{m \times 2d_h}$. These three output results correspond to distinct information sources or features. A one-dimensional maximum pooling operation is employed on these results to extract their most prominent features.

After the processing, three distinct pooled result vectors are generated. Initially, the first result corresponds to the max-pooling of sentence p_s. This implies that features with the highest values are selected for retention in sentences, thereby capturing the essential information. This process helps emphasize the most significant features in the sentences, enhancing its importance within the entire model. The second vector p_t corresponds to highlight the core information of the original topic. This contributes to ensuring that we maintain attention on the original topic throughout the entire processing and underscore its significance within the model. The output at last correlates to the max-pooling of the interaction topic px_t. The objective of this step is to identify the most salient features, and thus highlight the importance of the interaction theme within the model.

These vectors from three distinct information sources capture the most significant features from each information source, facilitating the model in better understanding and utilizing this information

for subsequent tasks. The concatenation of the three pooling exports yields the input vector $x \in R^{3*d_h}$ for the final module for categorization. The input to the final functional unit is the flattened concatenation outcome of the three pooling expressions from the previous actions, and the classification mechanism is identical to the first two models.

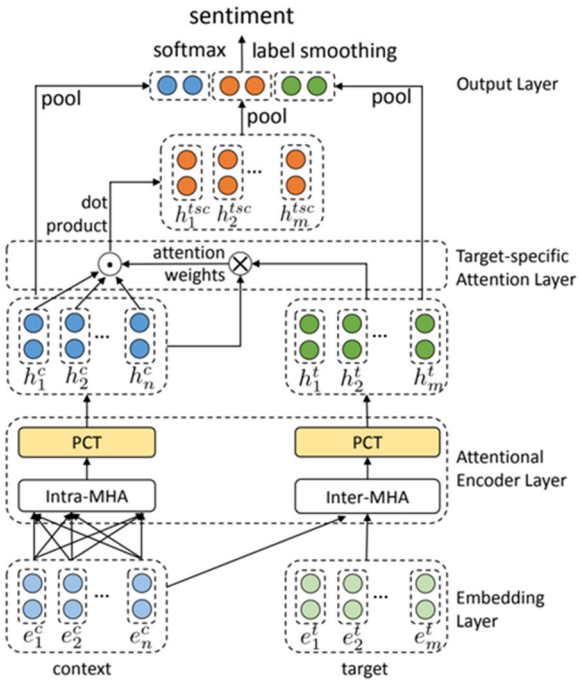

Fig. 3 The comprehensive structure of the AEN model.

3. Experimental design and analysis of results

3.1. Dataset

This article performs an initial comparative analysis using the publicly accessible SemEval 2014 Task 4 dataset provided by the International Workshop on Semantic Evaluation. The dataset comprises restaurant reviews (Restaurant) and laptop reviews (Laptop), with the detailed distribution presented in Table 1. The dataset is split into training sets (Train) and test sets (Test), both of which contain three emotional polarities: Positive, Neural, and Negative, corresponding to sentiment tendency values: 1, 0, and -1.

Table 1. Distribution of SemEval 2014 Task 4 dataset.

Dataset	Positive		Neural		Negative	
	Train	Test	Train	Test	Train	Test
Restaurant	2164	728	637	196	807	196
Laptop	994	341	464	169	870	128
Total	3158	1069	1101	365	1677	324

Utilizing the publicly available user-generated textual data on automobile-related content, provided by BDCI 2018, as the training set, the topics discussed in the comments as well as the sentiment information for each topic are analyzed and identified. This allows relevant person to gain insights into user preferences regarding the discussed topics. Discussion topics can be matched from the text or may require refinement based on context.

The training dataset comprises of 10,654 carefully annotated comments in total. The opening row with the words serves as the section header and is separated by semicolons. In this dataset, data id corresponds directly to content. However, a single text may encompass multiple topics, leading to multiple records for annotating different topics and sentiments. Consequently, duplicate data id values

are present across the entire training dataset. The test results are also formatted, with the first row serving as the header, and fields matching those in the training dataset. Following an 8:1:1 ratio, the training data gets separated into sets for training, validation, and testing.

Terminology annotation plays a crucial role in academic research by offering deeper understanding and annotation of research data or texts. Discovering comprehensive annotation details about fields aids in a further comprehension of the data. The description of training set fields of BDCI 2018 dataset provides an exhaustive description of the training set fields, delivering a thorough overview of data structures and features to ensure accurate comprehension and application of these fields. The name of the data ID field specifies its role as the unique identifier within the dataset, with an integer data type employed to ensure the uniqueness of identifiers for each data item. The textual content explicitly denotes its nature as containing text, with a data type of string indicating that it holds content information in text form. The analyzed topic clearly indicates its purpose in representing the subjects or domains related to textual content, with a string data type aiding in the categorization of textual content by topic. The score for sentiment signifies the presence of emotional analysis information. It is linked with an integer data type and is employed to describe the sentiment tendency or sentiment score of the analysed topic.

3.2. Evaluation Metric

This task uses the F_1 value rating index, which is based on the exact matching of theme + emotion words. T_p represents the number of correct judgments, F_p denotes the quantity of errors or overestimations, and F_n is a measure of the number of errors in judgment.

Then, the accuracy rate $P = \frac{T_p}{T_p+F_p}$, the recall rate $R = \frac{T_p}{T_p+F_n}$, and $F_1 = \frac{2*P*R}{P+R}$.

3.3. Model Comparisons and Analysis

3.3.1 Quantitative comparison on SemEval 2014 Task 4 dataset

Exhaustive statistics of the multiple types on the two datasets is shown in Table 2. Baseline results are taken from the original studies.

Table 2. Comparison results.

Models	Restaurant (Accuracy)	Laptop (Accuracy)	Mean Accuracy
IAN	0.7860	0.7210	0.7535
AOA	0.8120	0.7450	0.7785
AEN-GloVe	0.8098	0.7351	0.7725
AEN-BERT	0.8312	0.7993	0.8153

In addition, this paper compiled the dimensional variations between the models on Restaurant. All compared models were implemented with the same hyperparameters and infrastructure, and they were executed on the same GPU. Table 3 displays the memory usage of the models on the restaurant dataset. Models based on RNN (IAN) and those based on BERT (AEN-BERT) exhibit larger model dimensions.

Table 3. Model sizes.

Models	Model size	
	Params $\times 10^6$	Memory (MB)
IAN	2.16	15.30
AEN-BERT	112.93	451.84
AEN-GloVe	1.16	11.04

AEN-BERT demonstrates the highest performance in both of these domains. This suggests that the AEN model utilizing BERT as the embedding mechanism performs admirably in these two

domains. Nevertheless, its considerable model size may limit its practicality. AOA and AEN-GloVe exhibit slightly lower performance compared to AEN-BERT. AOA outperforms AEN-GloVe slightly, especially when considering the smaller memory footprint of the AEN model with GloVe embeddings. In cost-saving situations, AEN-GloVe emerges as the winner. Mean accuracy serves as a comprehensive assessment of these models' performance on the two datasets. This column reveals that the AEN-BERT model leads in terms of average performance. Among the other three models, IAN shows the smallest difference in mean accuracy compared to the accuracy for Restaurant and Laptop. This suggests that the performance of the IAN model is relatively stable, with minimal performance variation when dealing with sentiment analysis in different domains.

3.3.2 Application analysis on BDCI 2018

In order to compare the practical application results of different methods, we also conduct an additional set of experiments on BDCI 2018 dataset. The task data is segmented to acquire the dataset's lexicon and each word's insertion, and the resultant file is exported to the designated directory. As mentioned above, the set for training is divided into specified percentage for experimentation. For ease of execution, the parameters necessary for the model are set as defaults in the code, all are the best parameter settings after tuning on the validation set, and the project directory has been customized as needed so that different models can be run simply by typing python plus the model's name.

The cross-entropy loss for each subject is added to create the loss function that was used to train all three models. The model that received the highest F_1 score on the validation set is kept throughout the training phase and is then evaluated on the test set. Each model is also tested using multiple constant text lengths. Table 4 lists the assessment findings for the three models on the exam setting.

Table 4. Performance comparison for different fixed text lengths.

Models	Size of text	Accuracy	F_1
AOA	50	0.943	0.723
	150	0.941	0.724
	250	0.942	0.723
	350	0.941	0.722
IAN	50	0.941	0.715
	150	0.940	0.706
	250	0.939	0.702
	350	0.937	0.695
AEN	50	0.941	0.720
	150	0.942	0.722
	250	0.945	0.724
	350	0.943	0.723

Model performance varies as text length increases. The performance of the model decreases slightly as the text length increases. This may be due to the fact that longer texts contain more information and are more difficult to process, resulting in a decrease in performance.

The accuracy of AOA on different text lengths is relatively high, which indicates that AOA has a stable performance in handling sentiment analysis tasks with different text lengths.

AEN performs well on longer texts. This indicates that the AEN model is more suitable for processing longer texts because it is better able to capture more contextual information. IAN is slightly ahead on short texts. The IAN model is slightly ahead on shorter texts, but performance drops significantly on longer texts. This may indicate that the IAN model is more suitable for processing short texts, but less well suited for longer texts. In terms of F1 scores, both AOA and AEN show high scores on different text lengths, indicating that they can maintain high accuracy while maintaining good overall model performance. IAN has lower F1 scores across all text lengths, suggesting that despite good accuracy on shorter texts, the model has low recall, leading to a decrease in F1 scores.

In summary, AOA and AEN have exhibited excellent performance in this task, particularly when dealing with longer text lengths. IAN has a certain advantage on short texts but performs relatively poorly on long texts.

Based on the analysis above, we can select an appropriate model according to specific task requirements and the length of the text. When the task involves processing longer texts and cost efficiency is a concern, the AEN model is a viable option. For shorter text data, IAN or AOA can also be applicable. The optimal selection usually depends on the specific application scenarios and dataset characteristics. Choosing the appropriate model is crucial for sentiment analysis tasks with varying text lengths to maintain high performance and stability. Furthermore, the F1 score offers a more comprehensive assessment of model performance, especially when dealing with imbalanced data. It helps evaluate the trade-off between model precision and recall.

4. Conclusion

In this study, a comparative analysis of several classic end-to-end neural network models is performed to explore topic-level sentiment polarity analysis. For a specific task, specialized data is employed for data tokenization, model training, and evaluation. To perform experiments, the training data is separated by a certain ratio. The performance results of different models under the same evaluation metrics are visualized by evaluating the accuracy and F_1 score of each model at different fixed text lengths. At the end of the comparative analysis, a summary of the strengths, weaknesses, and performance in different aspects of each model is provided. The article discusses the existing problems of these models and puts forward suggestions for improvement, which can provide some new insights into the research of aspect sentiment analysis.

References

[1] LI Yang, WANG Shi,ZHU Junwu.Summarization of Aspect-level Sentiment Analysis [J]. Computer Science,2023,50(S1):34-40.

[2] HONG W, LI M.A Review of Research on Text Emotional Analysis Methods [J]. Computer Engineering and Science,2019,41(4):750-757.

[3] JOY H, KIM J, et al. GapFinder: Finding Inconsistency of Security Information from Unstructured Text[J]. IEEE Transactions on Information Forensics and Security,2020, PP (99):1-1.

[4] Huang B, Ou Y, Carley K M. Aspect level sentiment classification with attention-over-attention neural networks[C]. Social, Cultural, and Behavioral Modeling: 11th International Conference, Springer International Publishing, 2018: 197-206.

[5] Thet, Tun Thura, et al. Aspect-based sentiment analysis of movie reviews on discussion boards[J]. Journal of Information Science,2010,36(6);823-848.

[6] LIU B. Sentiment analysis and opinion mining [J]. Synthesis Lectures on Human Language Technologies,2012,5(1):1-167.

[7] Liu K, Xu L H. Sentiment analysis: mining opinions, sentiments and emotions[J]. Computational Linguistics,2016(3): 595-598.

[8] WANG Chundong, ZHANG Hui,MO Xiuliang,YANG Wenjun. Overview on sentiment analysis of microblog [J]. Computer Engineering & Science,2022,44(1):165-175.

[9] Tan Cuiping.Review of Fine-Grained Sentiment Analysis Based on Text[J]. Journal of Academic Libraries,2022,40(04):85-99+119.

[10] Nasukawa, Tetsuya, Jeonghee Yi. Sentiment analysis: Capturing favorability using natural language processing. *Proceedings of the 2nd international conference on Knowledge capture*. 2003.

[11] Jiang Long, et al. Target-dependent twitter sentiment classification. *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*. 2011.

[12] Tang Duyu, et al. Effective LSTMs for target-dependent sentiment classification. *arXiv preprint arXiv:1512.01100* (2015).

[13] Wang Yequan, et al. Attention-based LSTM for aspect-level sentiment classification. EMNLP. pp. 606-615. 2016.

[14] Ma Dehong, et al. Interactive attention networks for aspect-level sentiment classification. *arXiv preprint arXiv:1709.00893* (2017).

[15] Song Youwei, et al. Attentional Encoder Network for Targeted Sentiment Classification [J]. *arXiv preprint arXiv:1902.09314* (2019).