DSI DATA SCIENCE AND INFORMETRICS
VOL.3, NO.1, FEB. 2023

RESEARCH ARTICLE

Aspect–based sentiment analysis of online peer reviews and prediction of paper acceptance results

Minghui Meng[a], Ruxue Han[a], Jiangtao Zhong[a], Haomin Zhou[a], Chengzhi Zhang[a,b*]

a. Department of Information Management, Nanjing University of Science and Technology, Nanjing, China
b. Key Laboratory of Rich–media Knowledge Organization and Service of Digital Publishing Content Institute of Scientific & Technical Information of China, Beijing, China

ABSTRACT

Peer reviews of academic articles contain reviewers' overall impressions and specific comments on the contributed articles, which have a lot of sentimental information. By exploring the fine-grained sentiments in peer reviews, we can discover critical aspects of interest to the reviewers. The results can also assist editors and chairmen in making final decisions. However, current research on the aspects of peer reviews is coarse-grained, and mostly focuses on the overall evaluation of the review objects. Therefore, this paper constructs a multi-level fine-grained aspect set of peer reviews for further study. First, this paper uses the multi-level aspect extraction method to extract the aspects from peer reviews of ICLR conference papers. Comparative experiments confirm the validity of the method. Secondly, various Deep Learning models are used to classify aspects' sentiments automatically, with LCFS-BERT performing best. By calculating the correlation between sentimental scores of the review aspects and the acceptance result of papers, we can find the important aspects affecting acceptance. Finally, this paper predicts acceptance results of papers (accepted/rejected) according to the peer reviews. The optimal acceptance prediction model is XGboost, achieving a Macro_F_1 score of 87.43%.

KEYWORDS

Peer reviews; Aspect extraction; Sentiment analysis; Prediction of paper acceptance results

1 Introduction

Peer review is the reviewer's assessment of research results' competence, significance, and originality (Tennant, 2018; Kang et al., 2018). It is a necessary measure to ensure the quality of scientific information, and reduce errors and confusion, which plays a vital role in the scientific writing and publishing process (Wei et al., 2021). Open Peer Review (OPR) is an essential part of open science and has gradually developed in recent years. All aspects of the review process are made public in OPR. This increases the research's transparency and greatly enhances the fairness of the review process (Thelwall et al., 2020). With the development of OPR, many online peer review corpora on the Internet have gradually become available, pro-

* Corresponding Author: zhangcz@njust.edu.cn.

viding data support for studying online peer review comments. Peer review comments on academic papers are comment texts reviewed by domain experts on papers submitted in this field. They are the most common online peer review comments and can also be regarded as a particular type of comment. Peer review comments reflect the reviewers' overall impression of submitted papers and contain many specific details about the comment objects (Wang & Wan, 2018).

Currently, research on online peer review comments is mainly from the perspective of peer review process, text of the review comments, and scores of each review aspect. However, these researches only involve the superficial content of the online peer review comments and lack deeper mining. Review comments on the strengths and weaknesses of various aspects of papers contain much sentimental information. Mining the fine-grained emotion in online peer review comments is extremely important for both contributors and editors/chairs. For contributors, it provides insight into what aspects of the paper receive more attention during the review process, enabling them to revise and optimize the paper's content in a targeted manner. For editors/chairs, obtaining reviewers' emotional tendencies towards papers can provide them with additional decision-making perspectives for making final decisions.

Scholars have conducted related research on aspect-level sentiment analysis of review comments. But there are problems, such as the subjectivity of manually selected review aspects and review aspects of the overall evaluation of reviewers where the granularity is still too coarse. This paper aims to overcome these limitations by extracting the fine-grained aspects of academic papers' review comments and conducting a fine-grained aspect-level sentiment analysis. We explore the attention paid by reviewers to each aspect of papers and identify important aspects that impact the acceptance decision. The scores of review comments and the emotional distribution of review aspects are combined to predict the acceptance results of submitted papers automatically. This paper enriches the research on aspect extraction and sentiment analysis of online peer review comments by exploring the contents discussed above. That can also provide a reference for paper contributors to optimize their papers further and for paper decision-makers to make paper acceptance decisions.

2 Related Works

This section analyzes the current status, level, and development trend of online peer review comment mining and fine-grained sentiment analysis according to this study's research purpose and content.

2.1 Online peer review comment mining

Recently, peer review has attracted growing attention from scholars, leading to numerous studies on the corpus opening, text mining of peer review, and prediction of paper acceptance based on review comments. Kang et al. (2018) introduced the first open peer review comment dataset for research purposes, predicting the score of each aspect in a review based on the paper and review contents, and predicting the acceptance of a paper based on textual features. They found a high correlation between the overall and oral recommendation, and specific properties of a paper, such as having an appendix, correlate with a high acceptance rate. Hua et al. (2019) annotated and constructed the corpus AMPERE (Argument Mining for Peer Reviews) based on PeerRead, studied the content and structure of peer reviews under the argumentation mining framework by automatically detecting the argumentative propositions put forward by reviewers and their types (such as evaluating the work or

making suggestions for improvement). Wang and Wan (2018) proposed a multiple instance learning network with a novel abstract-based memory mechanism (MILAM) to address the task of automatically predicting the overall recommendation/ decision. It was found that there is generally good consistency between the review texts and the final recommended decisions, except for the borderline reviews. Ghosal et al. (2022) proposed a novel benchmark resource for computational analysis of peer reviews. They annotated the peer review report at the sentence level across four layers: Review-Paper Section Correspondence, Review-Paper Aspect, Review Statement Purpose, and Review Statement Significance. And they made a statistical analysis of the labels and their emotion distribution.

The continuous opening of online peer reviews provides the data support for us to conduct the content mining research of review comments. Mining review comments from the aspect level helps us to understand reviewers' intentions and emotional tendencies toward the object of evaluation more specifically. However, current aspect-based text mining research of peer review does not extract aspects from review comments, and directly adopts the reviewing aspects used in ACL conference, which are coarsely granular and lack more detailed comments. So this paper extracts the fine-grained aspects of review comments and explores the specific views and concerns of reviewers on each aspect of articles.

2.2 Fine–grained sentiment analysis

Peer review comments contain a variety of sentiments from reviewers about the aspects of papers. At present, sentiment analysis has gradually changed from coarse-grained sentiment analysis to fine-grained sentiment analysis. Scholars have done a lot of research on aspect extraction, aspect-based sentiment classification, and sentiment analysis of online peer review comments. Hu and Liu (2004) first proposed to extract the aspect words of goods by using the syntactic relations between sentiment words and aspect words. Experiments showed that this method is effective and universal in the English corpus. Qiu et al. (2011) proposed a double propagation (DP) algorithm, which utilizes syntactic relations to synchronously extract sentiment words and opinion objects (entities and aspects). Meng, Wang and Zhang (2021) conducted aspect extraction for peer review comments of academic papers for the first time. They proposed a multi-level aspect determination method and obtained a multi-level aspect set of peer review comments. Phan and Ogunbona (2020) used the shortest path between two words in the syntactic dependency parse tree as the syntactic relative distance (SRD) and proposed a new aspect-based sentiment classification model, LCFS-BERT, which performs best on the EMEVAL-2014 dataset. Ghosal et al. (2019a, 2019b) used the sentiment classifier, Vader, to annotate the sentiment polarity in peer review sentences, and added sentiment components to the deep neural structure to predict the paper's acceptance decision and aspect score. The results showed that adding sentiment information in review comments significantly improves the prediction ability of the system. Thelwall et al. (2020) proposed a sentiment analysis program, PeerJudge, to detect praise and criticism in peer review. PeerJudge is a dictionary based on the sentiment analysis method. The sentiment dictionary used by PeerJudge is composed of a manually encoded initial sentiment dictionary and machine learning adjusted and added sentiment dictionary. Chakraborty et al. (2020) proposed to use aspect-based sentiment analysis of scientific reviews to extract useful information. They found that the distribution of aspect-based sentiment obtained from a review is significantly different for accepted and rejected papers, and certain aspects present in a paper and discussed in the review strongly determine the final recommendation. Kumar et

al. (2022) proposed a novel deep neural architecture to make use of an aspect infused embedding. The experimental results showed that aspects, along with their corresponding sentiment, help to improve the performance of the peer review decision prediction system and assist the editor/chair in determining the outcome based on the reviews.

To sum up, the current mining of online peer review comments is mainly based on the review text or the review aspect itself. However the granularity of review aspects in the aspect-based sentiment analysis is coarse, these aspects pertain to an overall evaluation of the articles. Therefore this paper conducts the sentiment analysis of peer reviews at a fine-grained aspect level. By mining the aspect sentiment in the peer review comments, we can better understand the reviewers' attitude toward aspects, which can provide direction for submitters to optimize articles. This paper focuses on the fine-grained aspects and realizes the aspect-based sentiment automatic classification of peer review comments. Then we compare and analyze the sentiment distribution of online peer review comments of different acceptance results of papers. We mine the important aspects affecting the acceptance of papers and realize the prediction of paper acceptance results based on the sentiment distribution of review aspects.

3 Methodology

3.1 Research Framework

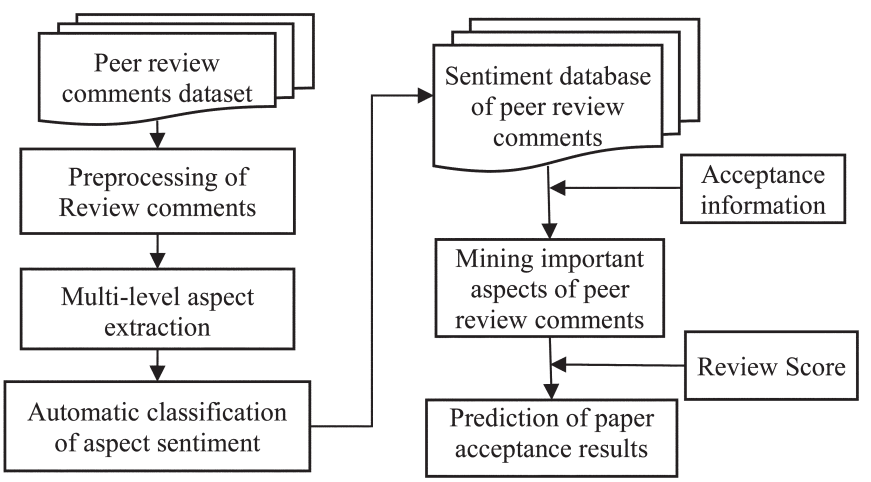

Figure 1 The framework of our work

This paper takes the online peer review comments of ICLR conference papers as an example, conducts aspect-level sentiment analysis of review comments and application research based on sentiment analysis. The research consists of four key components: multi-level aspect extraction of online peer review comments, automatic classification of aspect sentiment, important aspect mining, and paper acceptance prediction based on peer review comments. The research framework is shown in Figure 1.

3.2 Preprocessing of Review comments

First, preprocess the obtained online peer review comments, combine multiple review

comments for the same paper into a single txt file, and use Stanza[1] tool to segment sentences of the review corpus. Next, remove punctuation using regular expressions (r'[^\w\s]') and convert text to lowercase. Then, conduct word lemmatization. Finally, use the spaCy[2] to identify and extract noun phrases in the review corpus, and filter the first word with parts of speech "NN", "NNS", and "VBG" as the final noun phrase. Then join it with "_" as a new noun, and replace it in the original review corpus, which has satisfied the requirements of subsequent aspect extraction task.

3.3 Multi–level aspect extraction of online peer review comments

First, use the Double Propagation algorithm (DP) proposed by Qiu et al. (2011) to extract aspect words in review comments. This algorithm utilizes syntactic analysis to identify the dependency relationship between opinion words and evaluation objects. The sentiment words and aspect words are extracted simultaneously according to their linguistic and inter-word relationships. In this paper, we employ the Stanfordnlp[3] tool for POS labeling and sentences parsing of the corpora. This paper limits the opinion words' parts of speech to adjectives and the aspect words to nouns, and we restrict the dependency relations between opinion words and aspects to mod, subj, conj, etc. Then, the multi-level aspects in online peer review comments are determined using the method proposed by Meng et al. (2021). This method is based on the online peer review comments in the three-level discipline domains, determines the multi-level aspects by calculating the interdomain entropy (IDE) and termhood to measure the distribution uniformity and particularity of aspects across different domains. Finally, this paper employs the Affinity propagation (AP) algorithm (Frey & Dueck, 2007) to cluster the determined multi-level aspect words. Specifically, the CBOW model of Word2vec (Mikolov et al., 2013) is used to construct a 200-dimensional word vector for each aspect word. Then the cosine similarity matrix between the aspect words is input into the AP model, and adjust the preference value and damping coefficient. We use the Silhouette Coefficient (SC) (Peter, 1987) to evaluate the clustering performance and obtain peer review comments' final multi-level aspects clustering results. This method allows for more fine-grained aspects mining and divides the aspects into multiple levels to provide a collection of aspects from different perspectives. In addition, it reduces the uncertainty of manual screening and makes the aspect extraction results more reliable. Therefore, the aspects set of review comments obtained by the method used in this paper can better meet the mining needs of peer review comments.

3.4 Automatic classification of aspects–level sentiment in peer review comments

(1) Data annotation of reviews´ aspect–level sentiment

In this paper, sentiment annotation is performed on the extracted aspects set. The annotation specification is shown in Appendix A. The annotators are all NLP researchers familiar with the peer review process in the corresponding field.

Each piece of data in the corpus has the content of the review sentence, the corresponding opinion words, dependency relations and aspect words. The annotators mark the corresponding sentiment of each aspect in the review sentence. There are four kinds of sentiment

1 https://github.com/stanfordnlp/stanza

2 https://spacy.io/

3 http://nlp.stanford.edu/software/tagger.shtml.

polarity labels: positive (1), negative (-1), neutral (0), and fuzzy (2). The labeling format is shown in Table 1. To ensure the accuracy of the annotation, two annotators label each piece of data regarding the annotation specification. For the data with inconsistent labels, the third annotator manually check the annotation and choose one of the sentiment polarities marked by the two annotators or "fuzzy sentiment". To reduce the subjectivity of human judgment, annotators are required to label in strict accordance with the specifications. Finally, the "fuzzy sentiment" label data is filtered.

Table 1 Annotation format of aspects sentiment

Opinion words	Dependency relations	Aspect words	Label polarity	Review sentence
online	amod	method	0	*The authors propose an online purification method based on (clipped) iterative gradient ascent.*
limited	amod	patterns	–1	*the two datasets used in the paper represents limited visual patterns.*
important	amod	topic	1	*strengths: the pa per address the important topic of adversarial defence.*

(2) Construction of LCFS–BERT model for aspect sentiment classification

This paper chooses the pre-trained model LCFS-BERT proposed by Phan and Ogunbona (2020) to implement sentiment classification based on aspects in review comments. The LCFS-BERT model proposes Semantic Relative Distance (SRD) to analyze syntactic relationships between words to understand better the context related to target aspects. Figure 2 is the model diagram of LCFS-BERT classification, and the process of constructing the model is as follows.

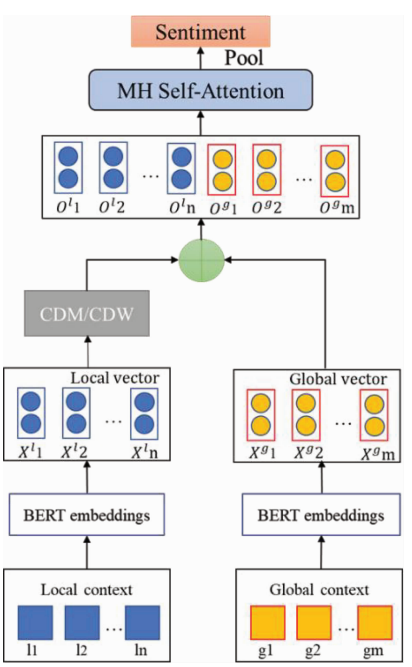

Figure 2 Model diagram of LCFS-BERT classification (Phan & Ogunbona, 2020)

Given {review sentence S, target aspect A}, the input layer processes the global context G as [CLS]+S+[SEP]+A+[SEP] and the local context L as [CLS] +S+[SEP]. The embedding layer uses two independent BERT word embedding models to encode them separately. A feature extractor in the feature selection layer, named Local Context Focus (LCFS), is designed to introduce the information of aspect words in the local context. LCFS uses the Context Dynamic Mask (CDM)/Context Feature Dynamic Weighting (CDW) mode to learn the global context feature. CDM masks low-semantics contextual feature whose semantic relevance to the target aspect words, as measured by SRD, is below a predefined threshold. These masked features are set as zero vectors. CDW reserves the contribution of contextual features with relatively few semantics but reduces their importance according to SRD between them and aspect words. Specifically, the SRD value between two words is calculated as the shortest distance between their corresponding nodes in the syntactic dependency tree. The output layer Average Pooling the encoded interaction feature representations, which are then fed into layers to predict from the set of sentiment polarity {positive sentiment, neutral sentiment, negative sentiment}. The entire model is fine-tuned using the cross-entropy and L_2 regulation as the loss function.

(3) Correlation Model of Aspect Sentiment Classification

In addition to the LCFS-BERT model, this paper also evaluates some sentiment classification models based on the BERT structure to implement the review aspect sentiment classification. The following are the basic principles of each model:

① **BERT method based on sentence–pair classification (BERT–SPEC) (Devlin et al., 2019)**

The BERT-SPC model based on the aspect sentiment classification task constructs the input sequence as "[CLS]" + global context + "[SEP]" + aspect word + "[SEP]". Followed by a fully connected layer and a Softmax layer, the probability results of each category are obtained, and the corresponding category results are obtained through the argmax operation.

② **Attention–encoding network based on targeted emotion classification (AEN–BERT) (Song et al., 2019)**

AEN-BERT is an attention encoding network that uses a pre-trained BERT model and an attention-based encoder to model context and target aspects. Then concatenate the vectors through average pooling, and the concatenated vectors are projected into the space of the target sentiment category using a fully connected layer.

③ **Local context focus model (LCF–BERT) (Zeng et al., 2019)**

The LCF-BERT model uses local context focus (LCF) and semantic relative distance (SRD) to accurately identify whether the context word is the local context of a specific aspect and discards words unrelated to the aspect word. The SRD value is calculated based on the distance between the two words. Based on SRD, the Context dynamic mask (CDM) and Context features Dynamic Weighted (CDW) are calculated to make the model focus on the local context. Additionally, multi-head self-attention is applied to simultaneously capture both the local and global contextual features of target aspects and fuse them together to predict the sentiment polarity of target aspects.

3.5 Important aspect mining of online peer review comments

In this paper, we select the optimal model of aspect sentiment classification to predict sentiment on the review corpus. We count the sentiment distribution and calculate the sentiment score of each aspect. The correlation between each aspect's sentiment score and the paper's acceptance result is calculated using the Spearman correlation coefficient. Through

the significance test method of the two-sided test, at the significant level of 0.01 and 0.05, we mine important review aspects related to acceptance. In this paper, the sentiment score of each aspect set of each article's review comments is defined as the sum of the average positive sentiment score, average neutral sentiment score and average negative sentiment score of the aspect set (Chakraborty et al., 2020). The formula is as follows:

$$Aspects_{sentscore} = \frac{(Pos_{score} + Neu_{score} + Neg_{score})}{review\_sentence\_num} \quad (1)$$

Among them, is the number of review sentences of the article, Pos_{score} is the positive sentiment score, Neu_{score} is the neutral sentiment score, Neg_{score} is the negative sentiment score, and $Aspects_{sentscore}$ is the sentiment score of the aspect set in the article.

3.6 Prediction of paper acceptance based on online peer review comments

This paper predicts the acceptance result of a paper (accepted/rejected) based on the distribution of review scores and aspect sentiment scores of review comments. The review score is an important part of the review comments, which plays a decisive role in the decision-making process for accepting a paper. In addition, the emotional tendencies of various aspects in review comments may affect the decision to accept papers, so we also add the emotional characteristics of review comments to the prediction model.

First, this paper calculates the average review score of each paper, obtains the characteristics of the review score and the characteristics of the aspect sentiment score of each paper's review comments. Then, the aspect features are further filtered using the correlation coefficient analysis and the important aspects are retained. Next, the characteristics of aspect sentiment and review scores are spliced horizontally to obtain the input data and standardize it with StandardScaler[4]. The prediction models, such as XGboost, Logistic regression, GRU, CNN are built to achieve the prediction of paper's acceptance result. Finally, use ten-fold cross-validation to train and evaluate the prediction classifier, and select evaluation indicators such as Accuracy, Precision, Recall, and F_1 value to obtain the best-suited classification model for the task. The following are the basic principles of each model:

(1) Logistic Regression (Genkin et al., 2007)

Logistic regression is a linear classification model. It has a linear decision boundary (hyperplane) and uses a nonlinear activation function (Sigmoid function) to simulate the posterior probability, so that the model output results between 0 and 1. The specific process is, given a data set $\{(x^{(1)},y^{(1)}),\ldots,(x^{(m)},y^{(m)})\}$, through Logistic regression, make the predicted value $\hat{y}^{(i)} \approx$ *actual value* $y^{(i)}$.

(2) Gradient Boosting Decision Tree (GBDT) (Friedman, 2001)

GBDT is an iterative decision tree algorithm. The algorithm consists of multiple weak classifiers. The weak classifier is a classification and regression tree (CART). Each iteration generates a weak classifier. Each classifier is trained on the residuals of the previous classifier, and the final strong classifier is obtained by a weighted summation of the classifiers trained in each round.

(3) XGboost (Chen & Guestrin, 2016)

XGboost is an optimization algorithm of GBDT. It uses the second-order gradient as an approximation of the residual. Since the loss function of GBDT does not take the complexity of the tree into account, XGBoost adds a regular term to penalize the complexity and improve the algorithm's generalization. Unlike GBDT, which uses the least-squares to calculate tree

4 https://scikit–learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

structure (CART), the loss function in XGBoost makes the result more accurate by doing a second-order Taylor expansion of the error.

The loss function of XGBoost adds a regularization term to the loss function of GBDT:

$$L_t = \sum_{i=1}^{m} L(y_i, f_{t-1}(xi) + ht(xi)) + \gamma J + \frac{\lambda}{2} \sum_{j=1}^{J} C_{tj}^2 \quad (2)$$

(4) Gated Recurrent Unit (GRU) (Chung et al., 2014)

GRU is a variant of LSTM. It is also a recurrent network. Compared with LSTM, its calculation is simpler and the amount of calculation is reduced. GRU has two gates, a reset gate and an update gate, which decide what information needs to be discarded or retained. This method uses a single GRU to model each input feature and obtains the hidden vector of each feature. These vectors are then fed into multiple fully connected layers for training, and the trained hidden vectors are finally input into the Logistic regression layer to predict the paper. The category of the admission result.

(5) Convolutional Neural Networks (CNN) (Kim, 2014)

Convolutional Neural Network consists of an embedding layer, a convolutional layer, a pooling layer, a fully connected layer, and an output layer. In the classification task, first, the feature is modeled by the embedding layer to obtain the vector representation of the feature. Then, the feature is extracted by the convolution layer. Next, the pooling layer samples the feature, the output dimension is reduced, and the important features are retained. Finally, connect multiple fully connected layers, train and update the feature vector, output the vector to the Logistic regression layer, and obtain the category of the paper's acceptance result.

4 Experiment and Result Analysis

This paper conducts experiments on the review comments of ICLR conference papers, covering multi-level aspect extraction of review aspects, automatic aspect sentiment classification of reviews, and sentiment analysis based on review comments. The following are the experiments and result analysis of each part.

4.1 Overview of the raw data of the ICLR review comments

Table 2 Data distribution of ICLR peer review comments

Acceptance results	Publication year	Number of papers	Number of reviews
Accepted	2017	172	588
	2018	422	1278
	2019	499	1519
	2020	681	2056
	2021	856	3297
Rejected	2017	255	856
	2018	534	1610
	2019	1058	3224
	2020	1870	5689
	2021	2118	8192
Total	2017–2021	8465	28309

Note: Collection date: March 10, 2021

ICLR (International Conference on Learning Representations) is one of the leading conferences in Machine Learning. It has a wide impact, a long time span and a large data scale on implementing the open review mechanism. In addition, it contains submitting papers on different acceptance results, which can meet multifaceted research. Therefore, this paper uses the paper review comments of the ICLR conference as the research data source and collects a total of 28309 open peer review comments from the ICLR from 2017 to 2021. The 2017 data is from the PeerRead dataset (Kang et al., 2018), and the 2018-2021 data is from the OpenReview website, as shown in Table 2.

4.2 Multi–level aspect extraction of ICLR paper review comments

This paper aims to mine the fine-grained aspects of review comments. Taking the review comments of ICLR papers as an example, a multi-level aspect extraction method is employed to identify common aspects unrelated to the field and special aspects related to the field.

(1) Candidate aspect extraction

Meng et al. (2021) determined the multi-level aspects extraction method of online peer review based on the online peer review comments of Nature Communications. It has three-level discipline fields, including the zero-level discipline field "NC", five first-level discipline fields and 71 second-level discipline fields. After investigation, ICLR belongs to the computer-related field, and the form of review comments is standardized. In this paper, the review comments of ICLR are classified under the second-level discipline "Mathematics and computing" of NC, extracting the multi-level candidate aspects of ICLR. The superior discipline field is "Physical science". A total of 8001 candidate aspects of the zero-level discipline, 5 groups of candidate aspects of first-level disciplines and 71 groups of candidate aspects of the second-level disciplines are extracted.

(2) Multi–level aspect determination

Table 3 Multi-level aspect cluster of ICLR review comments

Level	No.	Common aspects unrelated to the field	Level	No.	Common aspects related to the field	Level	No.	Special aspects related to the field
	0	Experiment & Result analysis		0	Organisms & Components		0	Algorithm
	1	Figures & Tables		1	Physical quantity		1	Performance
	2	Quantitative means & Operation		2	Reaction process		2	Training parameters
	3	Model & Method		3	Experimental operation		3	Learning & Training
	4	Technical indexes & Experimental parameters		4	Substance category		4	Experimental data
I	5	Function related	II	5	Substance structure	III	5	Optimizing strategies related
	6	Research conclusion & Discussion		6	Physical methods		6	Appendix & Theorems
	7	Language description		7	Research characteristics		7	Neural network structure
	8	Fields & Topics		8	Physical phenomenon		8	Inspection methods
	9	Research value		9	Viewpoint & Theme		9	Others
	10	Others		10	Others			

In this paper, the conditions for determining the multi-level aspect of ICLR are limited to the top 85% of the cumulative percentage of the total word frequency, having the entropy value higher than the median value, and selecting the top 200 of the term degree. And we filter the noise of non-aspect words that meet the above conditions (such as "other comment", "nature communication") and the candidate words of non-noun parts of speech (such as "support", "address"). After screening, we determine the multi-level aspect of ICLR, including 819 words of common aspects unrelated to the field, 575 words of common aspects related to the field, and 200 words of special aspects related to the field. To enhance both the clustering performance and distinctiveness in the multi-level aspect clustering results of ICLR, we set the number of clusters to 10-20 and the damping coefficient to 0.5-1. According to the meaning of words in various clusters, they are summarized into aspect categories. For common aspects unrelated to the field, the best contour coefficient is achieved with the damping coefficient is 0.6 and the preference value is -120. Two special clusters are manually merged into "other", and 11 clusters are obtained. For common aspects related to the field, the best contour coefficient is achieved with the damping coefficient is 0.55 and the preference value is -80. A non-aspect cluster is manually deleted, two clusters are classified into "Others" clusters, and 11 clusters are obtained. For special aspects related to the field, when the damping coefficient is 0.5 and the preference value is -20, a non-aspect cluster is manually deleted, two learning-related clusters are merged into "Learning & Training", and a total of 10 clusters are obtained. The multi-level aspect cluster of ICLR review comments obtained is presented in Table 3. See Appendix B for the detailed multi-level aspect set of ICLR.

(3) Comparison experiment of candidate aspect extraction

We use two methods to conduct the comparison experiment of aspect extraction with the multi-level aspect extraction method used in this paper: the aspect extraction method based on seed words (Mukherjee & Liu, 2012) and the aspect extraction method based on LDA (Blei et al., 2003). The aspect extraction method based on seed words (Mukherjee & Liu, 2012) was to obtain the top 1000 words in the frequency of review comments as seed words. It used word vectors to represent seed words and remaining candidate words, calculated the cosine similarity between two words, and selected the top 5 words most similar to the seed words as extensions. Then filtered non-words and obtained the final set of aspect words. The aspect extraction method based on LDA (Blei et al., 2003) used the LDA to model the topics of the ICLR review corpus. It generated 20 topics, then selected the first 100 words from each topic as candidate aspects and filtered non-words and non-noun noise to obtain the aspects set.

We randomly selected 20 papers accepted and 20 papers rejected from the ICLR conference. Three annotators who are familiar with the peer review process annotate aspect words of the review comments. We combine the different parts of speech of annotation results to obtain 807 aspects, which served as the gold standard for evaluating the extraction method. The results are evaluated using the precision (P-value), recall (R-value) and F_1 indexes. The formula is as follows. Where T is the number of aspect words correctly extracted, N is the number of aspect words incorrectly extracted, and U is the number of aspect words not extracted.

$$P = \frac{T}{T+N} \tag{3}$$

$$R = \frac{T}{T+U} \tag{4}$$

$$F_1 = \frac{2 \times P \times R}{P+\mathrm{R}} \tag{5}$$

Figure 3 shows the comparison of aspect extraction results of each method. As shown in the figure, the aspect extraction method used in this paper outperforms the other two methods in terms of P, R and F_1 values. It outperforms the LDA method by achieving higher P and R values, which may be because the LDA method is difficult to filter high-frequency non-aspect words and result in noise. However, our method can filter some non-aspect words through the distribution of candidate words in various fields to improve the extraction accuracy. In comparison to the seed words method, while the R-value is guaranteed through the expansion of the seed words, extracting aspects in remote fields is still limited by the choice of seed words. Therefore, our method has better adaptability for field aspects extraction.

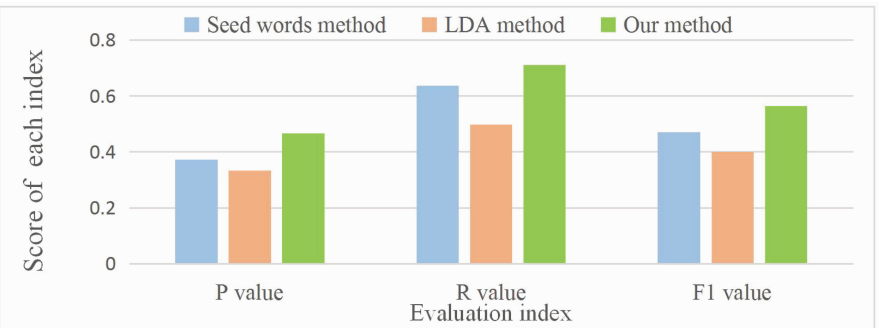

Figure 3 Comparison of aspect extraction results of each method

4.3 Automatic aspect sentiment classification of ICLR paper review comment

4.3.1 Overview of aspect sentiment classification dataset

We randomly selected the review comments of papers of different categories included in ICLR from 2017 to 2021 as the annotated corpus, including 20 papers accepted and 20 rejected. A total of 4390 sentences of reviews are annotated, including 2351 sentences of negative sentiment, 1295 sentences of neutral sentiment and 744 sentences of positive sentiment. To ensure the balanced distribution of the data set, the annotation data sets of accepted papers and rejected papers are separately divided into training and test sets in an 80/20 ratio and then combine them. The distribution of sentiment data sets is shown in Table 4.

Table 4 Distribution of sentiment data sets of ICLR

Sentiment category / Data category	Negative sentiment	Neutral sentiment	Positive sentiment
Training set	1658	1043	569
Test set	392	251	175

4.3.2 Experimental setup and evaluation method of aspect sentiment classification

(1) Experimental setup

For the LCFS-BERT model, both embedded dimension (bert_dim) and hidden dimension (hidden_size) are 768, the maximum sentence length (max_seq_len) is 80, learning rate (learning_rate) is set to 2e-05, attenuation rate (dropout) is set to 0.1, the regularization coefficient (L_2) is 0.01, the batch size (batch_size) is 16, the number of iterations (num_epoch) is 20, and

the semantic relative distance (SRD) is 3. Adam optimizer is selected for model optimization. The LCFS-BERT model code used in this paper is open source code[5], and the comparison model code is open source code[6]

(2) Evaluation methods

The aspect-level sentiment classification in this paper is a multi-classification task. We select indexes of Accuracy, Macro-Precision, Macro-Recall and Macro-F_1 to evaluate the performance of LCFS-BERT and each baseline model. The following is the calculation method of each index:

$$Accuracy = \frac{n_{correct}}{n_{total}} \tag{6}$$

$$Maro\_P = \frac{1}{n}\sum_{i=1}^{n} P_i \tag{7}$$

$$Macro\_R = \frac{1}{n}\sum_{i=1}^{n} R_i \tag{8}$$

$$Macro\_F_1 = \frac{2*Macro\_P*Macro\_R}{Macro\_P+Macro\_R} \tag{9}$$

Where, $n_{correct}$ is the number of correct classifications, n_{total} is the total number of categories and n is the number of categories.

4.3.3 Result analysis of aspect sentiment classification

Table 5 shows the experimental results of each aspect sentiment classification model on the ICLR peer review data set's test set. LCFS-BERT-CDM model achieves the best performance. The LCFS-BERT model uses the Context Dynamic Mask (CDM) method to extract context features, with an Accuracy value of 82.05% and a Macro-F_1 value of 80.56%. The LCFS-BERT model outperformed the BERT-SPEC and AEN-BERT models, which do not use additional knowledge from the specific corpus to train specific field embedding. This is likely due to the LCFS-BERT model implementing the CDM & CDW layers and paying more attention to the local context words of specific aspects by weakening the features with relatively fewer semantics. The Bert shared layer is more effective in extracting and learning context features (including local and global contexts). Both LCFS-BERT-CDM and LCFS-BERT-CDW models perform better than LCF-BERT-CDM and LCF-BERT-CDW models. It can be seen that the new semantic relative distance (SRD) proposed by the LCFS-BERT model shows improved understanding of target aspect and the context related to the target aspect by analyzing the syntactic relationship between words. And it significantly avoids the influence of long-distance context in the self-attention mechanism, so as to optimize its performance in this task, which is suitable for aspect sentiment classification of peer review comments.

Table 5 Results of each aspect sentiment classification model

Model	Accuracy(%)	Macro–P(%)	Macro–R(%)	Macro–F_1(%)
BERT–SPEC	81.56	81.23	78.49	79.46
AEN–BERT	80.71	80.56	78.89	78.55
LCF–BERT–CDW	81.07	79.24	78.47	78.55
LCF–BERT–CDM	81.20	82.43	76.44	78.39
LCFS–BERT–CDW	81.44	81.56	77.72	78.87
LCFS–BERT–CDM	**82.05**	81.16	80.06	**80.56**

5 https://github.com/HieuPhan33/LCFS–BERT

6 https://github.com/HieuPhan33/LCFS–BERT

4.4 Mining for important aspects of ICLR review comments

(1) Sentiment distribution of aspects

This paper chooses the LCFS-BERT model with the best performance to predict the sentiment of aspects of the unlabeled corpus. Examples of sentiment polarity prediction of aspects can be found in Appendix C.

Figures 4 and 5 show the heat maps of the sentiment distribution of the top 10 aspects in the accepted and rejected papers. The heat map shows the normalized frequencies of "Positive", "Neutral" and "Negative" sentiment for each aspect respectively, where deeper colors indicate higher frequencies. From the graph, it can be found that the popular aspects show a lower frequency of neutral sentiment. This may be explained by the fact that these aspects always carry the reviewers' sentiments as objects to be evaluated. Only a few aspects are mentioned merely or appear as part of the background introduction with a neutral sentiment. In addition, the frequency of negative sentiments is higher among top aspects. This is because the main content of the review comments is the reviewer's questions or suggestions about various aspects of the paper, and therefore the negative sentiment is more. In addition, in the review comments of accepted and rejected papers, aspects such as "method", "result", and "model" have the most emotions and are of great concern to the reviewers.

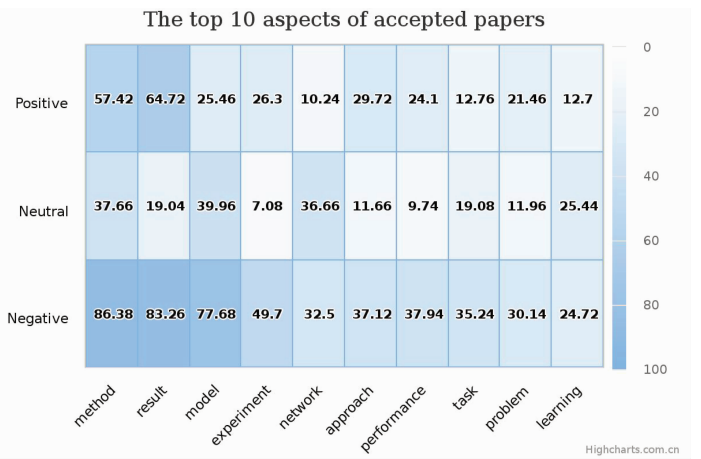

Figure 4 Heat map of sentiment distribution of top 10 aspects of ICLR accepted papers

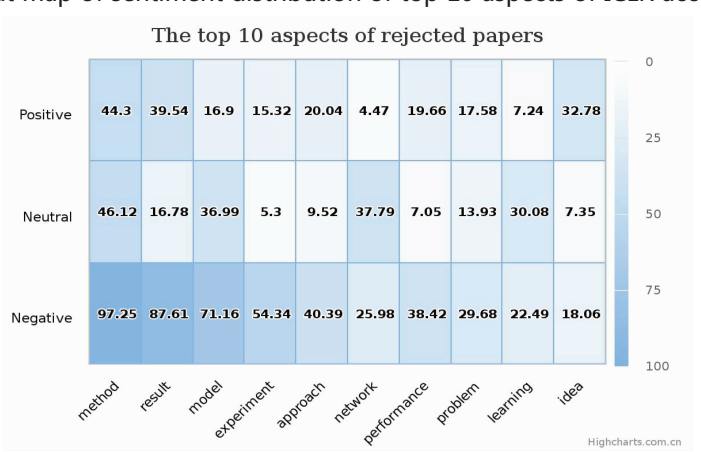

Figure 5 Heat map of sentiment distribution of top 10 aspects of ICLR rejected papers

(2) Distribution of review scores

In this paper, the review scores of accepted and rejected papers of ICLR 2017-2021 were counted separately and normalized. As shown in Figure 6, the horizontal axis represents the review score, and the vertical axis represents the rate of accepted (rejected) papers received that review score. The results reveal that a considerable number of accepted and rejected papers received the same review score. Thus, it is insufficient to determine the acceptance or rejection of a paper based on the review score alone. The review text is also an important criterion for making an acceptance decision.

Figure 6 ICLR 2017-2021 Distribution of review scores of accepted/rejected papers

(3) Correlation of review sentiment scores with paper acceptance result

Table 6 Correlation between sentiment score of paper review aspects and acceptance result

Level number		Aspect set	Spearman´s correlation coefficient	Level number		Aspect set	Spearman´s correlation coefficient
Ⅰ	0	Experiment & Result analysis	0.222**	Ⅱ	0	Organisms & Components	0.014
	1	Figures & Tables	0.038**		1	Physical quantity	–0.002
	2	Quantitative means & Operation	0.154**		2	Reaction process	0.007
	3	Model & Method	0.191**		3	Experimental operation	0.053**
	4	Technical indexes & Experimental parameters	0.066**		4	Substance category	0.024*
	5	Function related	0.060**		5	Substance structure	–0.017
	6	Research conclusion & Discussion	0.155**		6	Physical methods	0.024*
	7	Language description	0.117**		7	Research characteristics	0.004
	8	Fields & Topics	0.047**		8	Physical phenomenon	0.026*
	9	Research value	0.144**		9	Viewpoint & Theme	0.008
	10	Others	0.050**		10	Others	–0.009
Ⅲ	0	Algorithm	0.038**	Ⅲ	5	Optimizing strategies related	0.026*
	1	Performance	0.106**		6	Appendix & Theorems	–0.001
	2	Training parameters	0.021		7	Neural network structure	0.049**
	3	Learning & Training	0.055**		8	Inspection methods	0.101**
	4	Experimental data	0.031**		9	Others	0.002

Note: * indicates a significant correlation at the 0.05 level (two-tailed);
** indicates a significant correlation at the 0.01 level (two-tailed).

Table 6 shows the results of the correlation between the sentiment scores of the aspect sets of the peer review comments and the acceptance of papers at the ICLR conference. In the domain-independent common aspect set, all sets show significant correlation at the 0.01 level of significance. Specifically, the aspect sets of "Experiments & Results analysis" and "Model & Methods" have the highest relevance and are most valued by the reviewers. The second most relevant are the aspects of "Quantitative means & Operation", "Research conclusion & Discussion", and "Research value". The relevance of "Language description" is also higher than 0.1, which has a greater impact on the acceptance result. The correlation coefficients of "Figures & Tables" and "Technical indexes & Experimental parameters" are lower and less influence the acceptance results. For the common aspects related to the ICLR domain, only the set "Experimental operation" shows a significant correlation at the 0.01 level of significance and has the highest correlation with the acceptance results. At the 0.05 level of significance, the sentiment scores of the aspect set "Substance category", "Physics methods", and "Physical phenomenon" are significantly correlated with acceptance results, while all other aspects are not significant. The sentiment scores of most aspects did not correlate with their acceptance results. This may be because they are more closely related to the first-level disciplinary "Physical Science", thus having a limited impact on the acceptance of papers in ICLR. Finally, for the specific aspects related to the ICLR domain, the sentiment scores of the "Performance" and "Inspection methods" sets have the highest correlation with the acceptance results at the 0.01 level of significance. The aspects of "Algorithm", "Neural network structure", "Learning & Training", and "Experimental data" significantly correlate with the acceptance results. The "Optimizing strategies related" aspect set is significantly correlated at the 0.05 level of significance and has a small effect on the acceptance results. In addition, the aspect sets "Training parameters", "Appendix & Theorems", and "Others" are not significantly correlated at a 0.05 level of significance and are not correlated with the acceptance results.

4.5 Prediction of paper acceptance based on ICLR paper review comments

4.5.1 Sentiment dataset overview for ICLR´s review comments

This paper selected the ICLR conference review comments dataset from 2017-2021 to implement the paper acceptance prediction task. We excluded the 40 corpora manually labeled with aspect sentiment and kept only the corpora automatically labeled for aspect sentiment by the model. In this dataset, the aspect sentiment score of each paper's review comment is used as one piece of data, and there are a total of 8425 pieces of data. Table 7 shows the specific distribution of datasets with sentiment scores of ICLR aspects.

Table 7 Distribution of datasets with sentiment scores of ICLR aspects

Submission Year	Accepted Papers	Rejected Papers
2017	171	254
2018	418	527
2019	494	1057
2020	676	1869
2021	851	2108
2017–2021	2610	5815

4.5.2 Selection of hiring prediction characteristics

In this paper, for the 32 aspects of the ICLR conference paper review comments, each aspect is treated as a feature, with the aspect sentiment score serving as its feature value. Each paper is regarded as one piece of data, with a dimension of 1*32. We utilize the correlation coefficient method to further filter the aspect features based on the correlation between review sentiment scores and paper acceptance scores. Only the important aspects related to the paper acceptance results were retained, totaling 22 aspect sentiment features. In addition, we obtain the review score features with a dimension of 1*1. Finally, the review score features are horizontally spliced with the aspect sentiment score features with a dimension of 1*23 for each data.

4.5.3 Experimental setting and assessment methods for acceptance prediction

(1) Experimental setting

This task predicts the acceptance of a paper based on review comments, which is a binary classification task, with "1" represent to be accepted and "2" represent to be rejected. This paper builds a series of machine learning models using the scikit-learn[7] toolkit, including Logistic Regression, GBDT and Xgboost[8] models. In addition, a series of deep learning models, including CNN, GRU, etc., are built using the Keras[9] library. The experimental setup parameters of each model are shown in Appendix D. For the deep learning models, "Adam" was chosen as the optimizer, "Relu" as the fully-connected layer activation function, "sigmoid" as the output layer activation function, and "Adam" as the loss function, and we choose binary_crossentropy as the loss function. The learning rate is 0.0001, the epoch is set to 200, and the batch_size is 64. To prevent overfitting, a dropout layer with a decay rate of 0.2 is added between the model layers.

(2) Assessment methods

Based on the distribution pattern and size of the task dataset, this paper uses a hierarchical partitioning method (Stratified-K-Fold) to partition the data. Ten-fold cross-validation is also used to train and evaluate the predictive classifier. Table 8 shows the Confusion Matrix for the Binary classification problem of paper acceptance and rejection. Our selected evaluation metrics are Accuracy, Precision, Recall, F_1 value, Macro Accuracy (Macro_P), Macro Recall (Macro_R), and Macro Frequency Average (Macro_F_1). The specific calculation formulas are as follows.

Where $\text{Precision}_{\text{Acceptance}}$ is the accuracy of the model in the classification of the accepted paper category, $\text{Precision}_{\text{Rejection}}$ is the accuracy of the model in the classification of the rejected paper category, $\text{Recall}_{\text{Acceptance}}$ is the recall rate of the model in the classification of the accepted paper category, and $\text{Recall}_{\text{Rejection}}$ is the recall rate of the model in the classification of the rejected paper category.

Table 8 Confusion Matrix for the Binary classification problem of paper acceptance and rejection

Type \ Forecast Category	Positive Example	Negative Example
Positive Example	TP (Ture Positive Example)	FN (False Negative Example)
Negative Example	FP (False Positive Example)	TN (Ture Negative Example)

7 https://scikit–learn.org/
8 https://xgboost.readthedocs.io/en/latest/
9 https://keras.io/zh/

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{10}$$

$$Precision = \frac{(TP)}{(TP+FP)} \tag{11}$$

$$Recll = \frac{(TP)}{(TP+FN)} \tag{12}$$

$$F_1 = \frac{2TP}{(2TP+FP+FN)} \tag{13}$$

$$Macro\_P = \frac{Precision_{\text{Acceptance}}+Precision_{\text{Rejection}}}{2} \tag{14}$$

$$Macro\_R = \frac{Recall_{\text{Acceptance}}+Recall_{\text{Rejection}}}{2} \tag{15}$$

$$Macro\_F_1 = \frac{2*Macro\_P*Macro\_R}{Macro\_P+Macro\_R} \tag{16}$$

4.5.4 Analysis of admission prediction results

Table 9 shows the experimental results of each model. In general, the integrated model XGboost performs the best with an Accuracy value of 89.41% and a Macro_F_1 value of 87.43%. Its prediction F_1 value of 82.45% for accepted papers and 92.41% for rejected papers. This is all higher than other classification models. Logistic regression also performs well on this task. It has an Accuracy value of 89.15%, a Macro_F_1 of 87.24%, an F_1 value of 82.30% for accepted papers, and an F_1 value of 92.17% for rejected papers. It does not require scaling of the input features and is relatively suitable for Binary classification tasks. However, the weakness is that it is difficult to fit the real data distribution. Compared with several traditional machine learning models, the XGboost model improves the classification accuracy by combining several weakly supervised models to obtain a better and more comprehensive strongly supervised model. Compared with the integrated model GBDT, XGboost is optimized based on GBDT by adding a regular term to the cost function. This helps prevent overfitting and improves the generalization ability and accuracy of the algorithm, so the model classification performance is better than GBDT. The deep learning model does not perform as well as the XGboost model on this task may be due to the small amount of data for the experiment, the presence of noise and interference factors, and the overcomplicated network that can cause overfitting of the model. In addition, the structured data of this experiment is more suitable for machine learning model training, while the deep learning model has difficulty learning more features from it. In summary, the integrated model XGboost performs best in predicting paper acceptance results based on the sentiment scores of review aspects and can be used to predict the acceptance results of papers.

Table 9 Experimental results of each prediction model

Models Category	Model	Performance Indicators									
		Accepted + Rejected				Accepted			Rejected		
		Accuracy (%)	Macro_P (%)	Macro_R (%)	Macro_F_1(%)	Precision (%)	Recall (%)	F_1 (%)	Precision (%)	Recall (%)	F_1 (%)
Traditional machine learning-models	Logistic regression	89.15	87.46	87.06	87.24	83.10	81.57	82.30	91.81	92.54	92.17
	GBDT	89.02	87.52	86.53	86.99	83.82	80.00	81.84	91.22	93.06	92.13
	XGboost	**89.41**	88.05	86.93	**87.43**	84.67	80.42	**82.45**	91.43	93.44	**92.41**
Deep learning models	GRU	81.47	79.33	76.12	77.24	74.49	62.07	67.45	84.17	90.17	87.02
	CNN	86.98	85.32	83.72	84.42	81.38	75.17	78.10	89.26	92.27	90.73
	CNN+GRU	87.24	84.77	86.47	85.48	76.85	84.41	80.41	92.69	88.52	90.54
	CNN+LSTM	85.44	82.74	85.17	83.66	72.97	84.44	78.25	92.50	85.89	89.06

5 Conclusion and Future Works

This paper identifies a multi-level fine-grained aspect set of the peer reviews of ICLR conference papers. Compared to previous studies, a deeper level of mining is conducted. It is demonstrated through aspect extraction comparison experiments that the multi-level aspect extraction method proposed in this paper performs better. In the aspect-level sentiment classification task of reviews, the LCFS-BERT model performs best with the Accuracy of 82.05% and the Macro-F_1 of 80.56%. Then, this paper counts the sentiment distribution of review aspects and calculates the correlation between each aspect's sentiment score and the paper's acceptance result. Results show that "Experiment & Result analysis" has the highest correlation among the set of domain-independent common aspects. In the domain-specific aspects, "Performance" has the highest impact on acceptance, and "Appendix & Theorem" is not related to acceptance. Finally, this paper predicts the acceptance of papers based on the sentiment score of the review aspects and review score. The best model for predicting the acceptance of a paper is the XGboost, and the Macro_F_1 of ten-fold cross-validated achieves 87.43%.

There are still some weaknesses in the experiments of this paper. The extraction method of our review comment aspects is simpler. As well as the size of the paper acceptance prediction data is not large-scale enough. Additionally, only explicit sentiments in review comments are analyzed, etc. This will be an area for future improvement.

Acknowledgments

This work is supported by Opening fund of Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content (No. zd2022-10/02).

References

Blei, D. M., Ng, A. Y., & Jordan, M. I.(2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Chakraborty, S., Goyal, P., & Mukherjee, A.(2020). *Aspect–based sentiment analysis of scientific reviews*. Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, Wuhan, China, pp. 207–216.

Chen, T., & Guestrin, C.(2016). *Xgboost: A scalable tree boosting system*. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, San Francisco, USA, pp. 785–794.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y.(2014). *Empirical evaluation of gated recurrent neural networks on sequence modeling*. Proceedings of Nips 2014 Workshop on Deep Learning, Montreal, Canada, pp. 1–9.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *Bert: Pre–training of deep bidirectional transformers for language understanding*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, USA, pp. 4171–4186.

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science, 315*(5814), 972–976.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics, 29*, 1189–1232.

Genkin, A., Lewis, D. D., & Madigan, D. (2007). Large–scale Bayesian logistic regression for text categorization. *Technometrics, 49* (3), 291–304.

Ghosal, T., Kumar, S., Bharti, P. K., & Ekbal, A. (2022). Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *Plos one, 17* (1), e0259238.

Ghosal, T., Verma, R., Ekbal, A., & Bhattacharyya, P. (2019a). *A sentiment augmented deep architecture to predict peer review outcomes.* 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Champaign, USA, pp. 414–415.

Ghosal, T., Verma, R., Ekbal, A., & Bhattacharyya, P. (2019b). *DeepSentiPeer: Harnessing sentiment in review texts to recommend peer review decisions*. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp. 1120–1130.

Hu, M., & Liu, B. (2004). *Mining and summarizing customer reviews.* Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, USA, pp. 168–177.

Hua, X., Nikolov, M., Badugu, N., & Wang, L. (2019). *Argument mining for understanding peer reviews.* Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, USA, pp. 2131–2137.

Kang, D., Ammar, W., Mishra, B. D., Van Zuylen, M., Kohlmeier, S., Hovy, E., & Schwartz, R. (2018). *A dataset of peer reviews (PeerRead): Collection, insights and NLP applications.* 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, USA, pp. 1647–1661.

Kim, Y. (2014). *Convolutional neural networks for sentence classification.* Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, pp. 1746– 1751.

Kumar, S., Arora, H., Ghosal, T., & Ekbal, A. (2022). *DeepASPeer: towards an aspect–level sentiment controllable framework for decision prediction from academic peer reviews*. Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries, Cologne, Germany, pp. 1–11.

Meng, M., Wang, Y., & Zhang, C. (2021). *Building multi–level aspects of peer reviews for academic articles.* 18th International Conference on Scientometrics and Informetrics Conference (ISSI 2021), Online, pp. 1519–1520.

Mikolov, T., Chen, K., Corrado, G., et al. (2013). *Efficient estimation of word representations in vector space*. Proceedings of the International Conference on Learning Representations (ICLR 2013), Scottsdale, USA, pp. 1–12.

Mukherjee, A., & Liu, B. (2012). *Aspect extraction through semi–supervised modeling*. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju Island, Korea, pp. 339–348.

Peter, R. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational & Applied Mathematics, 20*, 53–65.

Phan, M. H., & Ogunbona, P. O. (2020). *Modelling context and syntactical features for aspect–based sentiment analysis.* Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, pp. 3211–3220.

Qiu, G., Liu, B., Bu, J., & Chen, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational linguistics, 37* (1), 9–27.

Song, Y., Wang, J., Jiang, T., Liu, Z., & Rao, Y. (2019). T*argeted sentiment classification with attentional encoder network.* International Conference on Artificial Neural Networks, Munich, Germany, pp. 93–103.

Tennant, J. P. (2018). The state of the art in peer review. *FEMS Microbiology letters, 365* (19), 1–10.

Thelwall, M., Papas, E. R., Nyakoojo, Z., Allen, L., & Weigert, V. (2020). Automatically detecting open academic review praise and criticism. *Online Information Review, 44* (5), 1057–1076.

Wang, K., & Wan, X. (2018). *Sentiment analysis of peer review texts for scholarly papers.* The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, USA, pp. 175–184.

Wei, C., Bu, Y., Kang, L., & Li, J. (2021). Directionality of paper reviewing and publishing of a scientist: A Granger causality inference. *Data Science and Informetrics, 1* (1), 68–80.

Zeng, B., Yang, H., Xu, R., Zhou, W., & Han, X. (2019). Lcf: A local context focus mechanism for aspect–based sentiment classification. *Applied Sciences, 9* (16), 3389.

Appendix

Appendix A: Aspect–level sentiment annotation specification for peer review comments

A.1 Background

Peer review comments on the academic paper are the text that experts review in journals or con–

ference papers in the field, including their overall impression of the paper and comments on the specific details of the paper. In order to discover the aspects that review experts pay attention to in the review process, aspect–level sentiment analysis is carried out for review comments. Aspects are the objects evaluated by reviewers in the review comments. By manually labeling the aspect sentiment of the review comments, a standardized annotation corpus is obtained to support the training of the sentiment classifier. This annotation standardizes the evaluation form of peer review corpora, combines the meaning of peer review sentences, finds the review aspects, and judges the emotional polarity of the review aspects. Through manual analysis, construct a peer review aspect–level emotional manual annotation system and standardize the annotation standards. Annotated corpora that meet this standard have high accuracy and strong normative, which is helpful for subsequent research.

A.2 Summary of review comments

① Describe the overall impression of the article, including the problem solved by the article, the article´s topic, the contribution of the article, etc.
② Comments on the details of the article, including methods, results, experiments, etc., mixed with factual descriptions and paraphrasing of the author´s views
③ Include job recommendations, requirements, assumptions, expectations
④ Include some clue words: figure, section, line, etc.
⑤ Suggest revisions to some minor mistakes in the article, such as typos, grammar, etc.
⑥ Consider the improvement or reduction of ratings, recommendation of articles, etc.
⑦ Evaluation of the author´s response, whether the issue was resolved, etc.

A.3 Judgment standard

A.3.1 Judgment standard of neutral aspect sentiment

① A direct description of the dissertation work, where the modifiers involved are neutral
E.g., The authors propose an online purification method based on iterative gradient ascent.
This paper studies adversarial defense by combing purification.
(Note: The words marked in the review sentence are aspects, the same below.)
② The modifiers for aspects are neutral, such as other, many, some, proposed, etc.
E.g., Figure 3 need to be better organized to group the results for the same dataset together
This architecture for online defense seems new (as far as I know).
③ Aspects that are only mentioned but not evaluated in the sentence are neutral, and the components of the sentence are neutral
E.g., The paper presents a novel method for answering "How many ... ?" questions in the VQA datasets.
The authors show results on How many–QA dataset for the proposed methods

A.3.2 Judgment standard of negative aspect sentiment

① Words that modify aspect words are derogatory words
E.g., The annotations and text in the figures are so small as to be almost unreadable (readability)
Table 2, the meaning of boldface is unclear.
② Include the reviewer´s expectations for some aspects of the paper, what they think can be improved, and how to do it better
E.g., To make the paper self–contained, it will be helpful to give more clarification on this.
It helps people to understand how strong this baseline is if you can confirm the implementation details in this part.
③ Sentences containing questions from reviewers about some aspect of the paper

E.g., Also, what are j_{1i} and j_{2i} in Theorem 2.1?

I´m confused by the claimed innovation in Lemma 3

④ The aspect word itself has negative emotions

E.g., typo, mismatch

⑤ Reviewers use derogatory terms to express their mood or emotions about an aspect

E.g., I am not sure the numbers reported for adversarial training match the state of the art reported in the mnist challenge leader.

⑥ Suggestions or requirements for a certain aspect of the article, such as: the sentence contains the words need, must, please, etc.

E.g., Figure 3 needs to be better organized to group the results for the same dataset together.

I encourage the authors to further refine the figures and writing to make this paper better.

⑦ There is a transition word before the clause where the aspect is located, and the previous sentence is a positive emotion

E.g., The paper is formally clear, **but** the discussion is **not** always at the same level of the technical ideas.

It is great, **but** my **concern** is generality of the method

⑧ Reviewers think that the authors did not address which issues they raised or which would be better if addressed, and the aspects involved in the question are negative

E.g., I will happily upgrade my rating of the paper if the authors can address my concerns over prior work in the experiments.

⑨ The overall score is low

E.g., Reasons for score, I vote for weak rejecting. In particular, I continue thinking that the contribution is limited. Accordingly, I did not change my scores.

A.3.3 Judgment standard of positive aspect sentiment

① Sentiment words that modify aspects are positive words

E.g., In my opinion these results are new and worth sharing.

The idea of zero–cost warmup and zero–cost move is very appealing.

② The aspect word itself has a positive sentiment

E.g., This paper demonstrates the effectiveness through extensive experiments.

The main contributions are the introduction of a unified framework that expresses 4 common attribution techniques

③ The sentence contains praise words to describe the paper, such as cost reduction, time reduction and performance improvement, etc., which need to be analyzed according to the actual situation

E.g., The idea is **novel** and **sound**.

Only using zero–cost proxies might not achieve competitive performance, but this idea provides a practical way of using these zero–cost proxies that will yield nice performance and a great reduction in search cost.

④ Reviewers use compliments when expressing their mood or emotions about aspects

E.g., I like the idea of zero–cost proxies which decides the performance of neural architectures at initialization and the high uyility of the zero–cost proxies available combining it with various NAS methods.

I really enjoy the derivation of the beta–elbo in the zero limit

⑤ Sentences contain different emotional information, which are positive emotions as a whole; that is, positive emotions are included after transition sentences

E.g., The architecture does not look entirely novel, **but** I kind of **like** the simple and practical approach compared to prior work.

Only using zero–cost proxies might not achieve competitive performance, but this idea provides

a practical way of using these zero–cost proxies that will yield nice performance and a great reduc–tion in search cost.

⑥ Although some emotional words do not appear together with aspect words, they also express e–motional expressions for a certain aspect (syntax relation xcomp). The common ones are:

vague / obscure: ´clarity´ –1 , clearly: ´clarity´ +1,

first time: ´innovation´ +1,

easy read / easy follow / easy understand: ´readability′ +1,

well described: ´description´ +1, well written: ´writing′ +1

⑦ The reviewer will increase the review score due to certain aspects

E.g., Given, that the authors were able to improve the results in the sequential MNIST and improve the average baselines, my rating improves one point.

A.3.4 Judgment Standard of Modal Verb

① Can / could: expressing the ability to accomplish something, positive emotion

② Can / could + Comparatives: negative emotions

③ Could / would + Positive emotion words: negative emotion

④ Could / would + Comparatives: negative emotions

⑤ Should / shall + Comparatives: negative emotion

⑥ Should / shall + Negative words: negative emotion

⑦ Should + do / suggest / please / encourage / recommend: negative emotion

⑧ Need + Negative words: positive emotion; Need + Comparatives: negative emotion; Need: negative emotion

⑨ Must+ Negative words: positive emotion; Must+ Comparatives: negative emotion

⑩ Have to / had better / better + Negative words: positive emotion

A.4 Labeling Rules

For the convenience of labeling, this paper first divides the target text into sentences according to the clauses ".", "?", "! " and "...". Then, divide the sentence into clauses according to "," and import the clauses into Excel. The aspect words are extracted by using a series of rules such as the rela–tionship between aspect words and opinion words, and the (opinion words, syntactic relationship, aspect words, clauses) are stored in the excel file in rows, and the aspect words are marked in the sentence corresponding emotion.

A.4.1 Aspect sentiment annotation of peer review

The annotations are divided into four categories: positive emotion, negative emotion, neutral emo–tion, and fuzzy emotion. The specific labeling method is shown in Table A1.

Table A1 Peer review aspects annotation

Sentiment classification of aspects	Definition	Label
negative sentiment aspect	The sentiment contained in this aspect is derogatory	–1
neutral sentiment aspect	The sentiment contained in this aspect is neutral	0
positive sentiment aspect	The sentiment contained in the aspect is positive	1
fuzzy sentiment aspect	Unable to determine the sentiment of the aspect or determine that the word is not an aspect	2

A.4.2 Annotation Format

Table A2 is the annotation format in Excel.

Table A2 Annotation Format

Opinion words	Syntactic relationship	Aspect words	Label polarity	Review sentence
online	amod	method	0	*The authors propose an online purification method based on (clipped) iterative gradient ascent.*
limited	amod	patterns	–1	*the two datasets used in the paper represents limited visual patterns.*
important	amod	topic	1	*strengths: the paper address the important topic of adversarial defence.*

A.5 Notes

① If there are multiple aspects in a sentence, mark the sentiment polarity of each aspect separately.
② First, determine whether the sentence has an emotional tendency, and then mark the emotional polarity of the aspect.
③ Only judge the sentiment polarity of the aspect words extracted from the table.
④ If there is no transition word, then the emotional polarity of the aspects involved before and after the sentence is the same.
⑤ Entity and aspect emotional tendencies are the same, such as "the result of experiment", where the polarity of "result" and "experiment" is the same.
⑥ If an aspect in a sentence is repeatedly extracted, if the aspect appears multiple times in the sentence and there is no obvious emotional transition word, the polarity is the same; if it appears only once, the second mark is 2.
⑦ For the advantages or disadvantages mentioned in the sentence, the emotional polarity of the aspect of the sentence corresponds to it.

A.6 Annotation Process

During the labeling process, the sentiments of the aspect words contained in each review sentence are labeled by three annotators who are all NLP researchers and are familiar with the peer review process in the corresponding field. In the annotation framework of this paper, each annotator must annotate the corresponding sentiment for the aspect existing in the review sentence, and the sentiment polarity corresponding to each aspect is one of four labels-"positive", "negative", "neutral" and "fuzzy". First, this paper uses the Stanfordnlp tool to analyze the review sentences syntactically, and uses the syntactic relationship between opinion words and aspect words to extract the aspect words existing in the review sentence. The syntactic relationship includes amod, subj, conj, etc. and personnel to mark the corresponding sentiment of the aspects in the review sentence. Due to the difficulty of labeling, to ensure the quality of the labeling, this labeling work does not carry out a labeling consistency check, but the third person manually checks the labeling data of the two people, and the labels are inconsistent. For the data, select one of the emotional polarities marked by the two people or mark the emotional polarities as "fuzzy". Finally, perform statistics on the marked data. It should be noted that a particular sentence may have no aspects or sentiments.

Appendix B: Multi–level aspect set of ICLR

Level	No.	Aspects set	Top5 Aspect words				
I	0	Experiment & Result analysis	data	result	experiment	example	information
	1	Figures & Tables	figure	text	table	sentence	image
	2	Quantitative means & Operation	analysis	development	evaluation	validation	characterization
	3	Model & Method	model	method	approach	mechanism	system
	4	Technical indexes & Experimental parameters	number	time	value	sample	condition
	5	Function related	effect	function	fact	factor	feature
	6	Research conclusion & Discussion	discussion	question	issue	conclusion	detail
	7	Language description	find	addition	claim	lack	understand
	8	Fields & Topics	structure	region	site	range	area
	9	Research value	role	interest	impact	quality	novelty
	10	Others	level	activity	loss	target	treatment
II	0	Organisms & Components	cell	membrane	animal	marker	plasma
	1	Physical quantity	temperature	absorption	compression	decomposition	conductivity
	2	Reaction process	activation	expression	pathway	localization	disease
	3	Experimental operation	setup	operation	spectroscopy	dft calculation	detector
	4	Substance category	water	metal	carbon	emission	triplet
	5	Substance structure	object	atom	symmetry	lattice	beam
	6	Physical methods	tune	compute	capture	grasp	showcase
	7	Research characteristics	originality	scalability	simplicity	demand	essence
	8	Physical phenomenon	assembly	realization	retrieval	execution	supply
	9	Viewpoint & Theme	viewpoint	suite	literature review	guidance	sketch
	10	Others	subsection	footnote	fee	subscript	missing reference
III	0	Algorithm	algorithm	computation	existing method	learning method	baseline method
	1	Performance	baseline	performance	generalization	effectiveness	test accuracy
	2	Training parameters	update	iteration	batch	initialization	epoch
	3	Learning & Training	optimization	reward	reinforcement learning	reinforcement	attack
	4	Experimental data	training data	image classification	data augmentation	experiment result	benchmark datasets
	5	Optimizing strategies related	gradient	regularization	guarantee	gradient descent	norm
	6	Appendix & Theorems	appendix	theorem	notation	lemma	proposition
	7	Neural network structure	task	network	architecture	representation	latent
	8	Inspection methods	ablation	motivation	ablation study	benchmark	variant
	9	Others	machine	vision	segmentation	computer vision	machine translation

Note: Since too many aspect words are in each aspect set in multi-level aspects of ICLR, it is impossible to display them all. So this appendix only shows each aspect set's first five aspect words.

Appendix C: Examples of sentiment polarity prediction for different aspects

Level	No.	Aspects set	Aspects	Sentiment Polarity	Review sentences
I	0	Experiment & Result analysis	experiments	–1	*do the authors have any insights or experiments on how looser relaxations, which would lead to less feature available would fare?*
	1	Figures & Tables	figures	1	*well –written paper with clear figures and explanations.*
	2	Quantitative means & Operation	verification	0	*this paper deals with complete formal verification of neural network, based on the branch and bound framework.*
	3	Model & Method	method	1	*as the method is interesting and analysis is quite thorough it´s easy for me to recommend acceptance.*
	4	Technical indexes & Experimental parameters	node	0	*the description of the nodes indicates that all hidden activation have a representative node in the gnn.*
	5	Function related	effects	–1	*more detail on the adaptive coder and its effects should be provided, and I will be happy to give a higher score when the authors do.*
	6	Research conclusion & Discussion	topic	1	*the paper is on a highly–relevant topic and explores a useful practical trick.*
	7	Language description	claim	–1	*I am quite surprised and not sure if this claim is true.*
	8	Fields & Topics	branching	–1	*how accurate is the learned heuristic in imitating strong branching?*
	9	Research value	novelty	–1	*cons: – novelty is somewhat low, as it is a straightforward application of existing ideas like gasse et al.*
	10	Others	formation	0	*review: authors describe a procedure of building their production recommender system from scratch, begining with formulating the recommendation problem, label data formation, model construction and learning.*
II	0	Organisms & Components	shot	0	*this paper argues about limitations of rnns to learn models than exhibit a human–like compositional operation that facilitates generalization to unseen data, ex. zero–shot or one–shot applications.*
	1	Physical quantity	digits	0	*for example, a larger input domain (as one of the reviewers also mentions) is mnist digits and we can imagine a problem where the npi must infer how to sort mnist digits from highest to lowest.*
	2	Reaction process	fusion	–1	*also, need to compare with this type of shallow fusion.*
	3	Experimental operation	operation	0	*this paper argues about limitations of rnns to learn models than exhibit a human–like compositional operation that facilitates generalization to unseen data, ex. zero–shot or one–shot applications.*

Level	No.	Aspects set	Aspects	Sentiment Polarity	Review sentences
Ⅱ	4	Substance category	water	–1	*the empirical comparison with tbptt is substantial but the water are muddied a bit by imprecise_presentation of baseline*
	5	Substance structure	vertices	0	*it seems the gnn has maNY vertices – the same number as the number of neurons in a network, which can be quite large.*
	6	Physical methods	gauge	–1	*for the claim that the algorithm does better, this is also difficult to gauge because the graphs are unclear.*
	7	Research characteristics	scalability	1	*k and pcc@k, and also with good scalability.*
	8	Physical phenomenon	completion	0	*the learnt model is then used to perform tree–beam search using a search algorithm that searches for different completion of trees based on node types.*
	9	Viewpoint & Theme	viewpoint	0	*the authors did not compare with existing work that tries to improve the robustness of neural nets from a differential equation viewpoint.*
	10	Others	subfigures	–1	*given that many figures have several subfigures, the authors should consider using a package that will denote subfigures with letters.*
Ⅲ	0	Algorithm	trick	1	*the paper is on a highly–relevant topic and explores a useful practical trick*
	1	Performance	performance	1	*the authors also discuss fallback mechanism to prevent bad failures case, as well as an online fine–tuning strategy that provide better performance.*
	2	Training parameters	training_time	–1	*the authors should make sure to include the ablation_study results, and a detailed discussion on training_data generation time and training_time in the final version of the paper.*
	3	Learning & Training	training	1	*in addition to showing the efficacy of ´deep learning´ for a new application, a key contribution of the paper is the introduction of a differentiable version of "rate" function, which the authors show can be used for effective training with different rate–distortion trade–offs.*
	4	Experimental data	training_data	–1	*the authors should make sure to include the ablation_study results, and a detailed discussion on training_data generation time and training_time in the final version of the paper.*
	5	Optimizing strategies related	quantization	–1	*while the results show that the method can work in theory with <8–bit activations, I am not sure how the quantization scheme could be efficiently implemented on actual hardware.*
	6	Appendix & Theorems	notation	1	*I also appreciate that the same notation was used in this paper and the original deep mind paper*
	7	Neural network structure	networks	0	*the tactic employed here is to learn a graph neural network (which allows to transfer the heuristic from small networks to large networks), using supervised training to imitate strong branching.*
	8	Inspection methods	ablation_study	–1	*the authors should make sure to include the ablation_study results, and a detailed discussion on training_data generation time and training_time in the final version of the paper.*
	9	Others	vision	–1	*I think the paper could be improved immensely by some empirical analysis of the rank of compressed standard vision networks and rank of activation covariance matrices.*

Appendix D: Experiment parameters of each acceptance prediction model

Model	Parameters
Logistic Regression	*penalty='l2'; C=1e5*
GBDT	*n_estimators=100*
Xgboost	*learning_rate=0.0001, n_estimators=100, max_depth=2*
GRU	*GRU(256) +Dense(128)+Dense(1)*
CNN	*Conv1D (256,3)+ MaxPooling1D (3,3)+ Conv1D (128,3)+ MaxPooling1D (3,3)+ Conv1D (64,3)+ MaxPooling1D(3,3)+ Flatten()+Dense(128)+Dense(1)*
CNN+GRU	*Conv1D (256,3)+ MaxPooling1D (3,3)+ Conv1D (128,3)+ MaxPooling1D (3,3)+ Conv1D (64,3)+ MaxPooling1D(3,3) +GRU(256)+ Dense(256)+Dense(128)+Dense(1)*
CNN+LSTM	*Conv1D (256,3)+ MaxPooling1D (3,3)+ Conv1D (128,3)+ MaxPooling1D (3,3)+ Conv1D (64,3)+ MaxPooling1D(3,3)+LSTM(256)+ Dense(256)+ Dense(128)+Dense(1)*