Hierarchical Bi-Directional Self-Attention Networks for Paper Review Rating Recommendation

Zhongfen Deng[1], Hao Peng[2,3], Congying Xia[1] , Jianxin Li[2], Lifang He[4] , Philip S. Yu[1]
[1]Department of Computer Science, University of Illinois at Chicago, Chicago, USA
[2]BDBC, Beihang University, Beijing, China
[3]School of Cyber Science and Technology, Beihang University, Beijing, China
[4]Department of Computer Science and Engineering, Lehigh University, Bethlehem, USA
{zdeng21,cxia8,psyu}@uic.edu, {penghao,lijx}@act.buaa.edu.cn
lih319@lehigh.edu

Abstract

Review rating prediction of text reviews is a rapidly growing technology with a wide range of applications in natural language processing. However, most existing methods either use hand-crafted features or learn features using deep learning with simple text corpus as input for review rating prediction, ignoring the hierarchies among data. In this paper, we propose a **H**ier**a**rchical **b**i-directional self-attention **Net**work framework (HabNet) for paper review rating prediction and recommendation, which can serve as an effective decision-making tool for the academic paper review process. Specifically, we leverage the hierarchical structure of the paper reviews with three levels of encoders: sentence encoder (level one), intra-review encoder (level two) and inter-review encoder (level three). Each encoder first derives contextual representation of each level, then generates a higher-level representation, and after the learning process, we are able to identify useful predictors to make the final acceptance decision, as well as to help discover the inconsistency between numerical review ratings and text sentiment conveyed by reviewers. Furthermore, we introduce two new metrics to evaluate models in data imbalance situations. Extensive experiments on a publicly available dataset (PeerRead) and our own collected dataset (OpenReview) demonstrate the superiority of the proposed approach compared with state-of-the-art methods.

1 Introduction

With an increasing submission of academic papers in recent years, the task of making final decisions manually incurs significant overheads to the program chairs, it is desirable to automate the process. In this study, we aim at utilizing document-level semantic analysis for paper review rating prediction and recommendation. Given the reviews of each paper from several reviewers as input, our goal is to infer the final acceptance decision for that paper and the reviewers' evaluation with respect to a numeric rating (e.g., 1-10 points). Paper review rating prediction and recommendation is a practical and important task in AI applications which will help improve the efficiency of the paper review process. It is also intended to enhance the consistency of the assessment procedures and outcomes, and to diversify the paper review process by comparing human recommended rating with machine recommended rating. In the literature, most of existing studies cast review rating prediction as a multi-class classification/regression task (Pang and Lee, 2005). They build a predictor by using supervised machine learning models with review texts and corresponding ratings. Due to the importance of features, most researches focus on extracting effective features such as context-level features (Qu et al., 2010) and user features (Gao et al., 2013) to boost prediction performance. However, feature engineering is time-consuming and labor-intensive.

Recently, various deep learning-based models have been proposed for automatically learning features from text data (Bengio et al., 2013). Existing deep learning models usually learn continuous representations of different grains (e.g., word, phrase, sentence, document) from text corpus (Pennington et al., 2014; Lai et al., 2015; Kim, 2014; Conneau et al., 2017; Wang, 2018; Qiao et al., 2018). Although deep learning models can automatically learn extensive feature representation, they cannot efficiently capture the hierarchical relationship inherent to the review data. To address this problem, Yang et al.

Proceedings of the 28th International Conference on Computational Linguistics, pages 6302–6314
Barcelona, Spain (Online), December 8-13, 2020

(2016) studied a hierarchical architecture and implemented it in deep learning framework to learn a better document-level representation. Also, with the success of attention mechanism in many tasks such as machine translation, question answering and so on (Vaswani et al., 2017), Shen et al. (2018a) designed a directional self-attention network to gain context-aware embeddings for words and sentences. Despite great progress made by these models, they do not focus on the task of paper review rating recommendation and are not effective enough to be directly used for this task because of the following reasons: First, the review data is hierarchical in nature. There exists a three-level hierarchical structure in the review data: word level, intra-review level and inter-review level, while previous models only capture two-levels (i.e., the word level and intra-review level) of this hierarchy. Second, paper reviews are usually much longer than other reviews (e.g., product reviews, movie reviews, restaurant reviews, etc.), while most of these models are working on those shorter reviews stated above and they do not leverage the up to date representation techniques such as BERT (Devlin et al., 2019) and SciBERT (Beltagy et al., 2019).

In this paper, we propose a novel neural network framework for paper review rating recommendation by taking word, intra-review and inter-review information into account. Specifically, inspired by HAN (Yang et al., 2016) and DiSAN (Shen et al., 2018a), we introduce a Hierarchical Bi-directional self-Attention Network (HabNet) framework to effectively incorporate different levels of hierarchical information. The proposed framework consists of three main modules in end-to-end relationship: sentence encoder, intra-review encoder and inter-review encoder, which can consider hierarchical structures of review data as comprehensive as possible. The outputs of inter-review encoder are leveraged as features to build the rating predictor without any feature engineering. We release the code and data collected by us to enable replication and application to new tasks, available at *https://github.com/RingBDStack/HabNet*.

The contributions of this work are as follows:

- We present a novel framework to guide the investigation and assessment of the effects of hierarchies on review data. To our best knowledge, this is the first work incorporating different levels of semantic information into a hierarchical neural network to perform paper review rating recommendation.
- We introduce two new metrics to better evaluate models when the distributions of classes are highly imbalanced (such as the paper review data we are working with).
- Empirical results on OpenReview (ours) and extended PeerRead datasets demonstrates the effectiveness of the proposed method in automatically making final acceptance decisions and helping reveal the rating inconsistency between the semantic review content and the numerical review ratings.

2 Related Work

2.1 Review Rating Prediction

Review rating prediction is a basic task in sentiment analysis. It was initially studied by Pang and Lee (2005) who cast this problem as a multi-class classification/regression task. In the literature, most of studies following this approach used supervised machine learning models to do review rating prediction. Since the features used by these models are critical for prediction performance, more refined textual features are exploited. Qu et al. (2010) introduced bag of opinions representation, where an opinion was composed of a root word, a set of modifier words and one or more negation words. Gao et al. (2013) used user-specific and product-specific features to increase the reliability of sentiment classification. With the popularity of deep learning model, instead of hand-crafted features, many works were proposed to automatically learn features from text corpora. Lai et al. (2015) applied a recurrent structure for convolutional neural network to capture contextual information for learning word representation. Conneau et al. (2017) used very deep convolutional networks to learn hierarchical representations of whole sentences. Johnson and Zhang (2017) studied deepening word-level CNNs to capture global representations of text. Peng et al. (2018) designed a deep Graph-CNN to learn both non-consecutive and long-distance features of text.

Kang et al. (2018) collected a dataset of peer reviews from several conferences and predicted paper acceptance decision by using paper draft. Gao et al. (2019) focused on predicting after-rebuttal scores by using their presented corpus. Hua et al. (2019) applied argument mining on their AMPERE dataset

to assess the efficiency of reviewing process. Li et al. (2019) designed a neural model to predict citation count of accepted papers. Yang et al. (2018) designed a hierarchical attention-based CNN for automatic academic paper rating by using source paper, it adopts original attention mechanism which cannot capture the interactions between elements in the same level. Leng et al. (2019) proposed DeepReviewer for automatic paper review utilizing paper's grammar and innovation to help learn better representation and predict paper's final review score. Different from above works, we aim at predicting the final acceptance decisions for papers and ratings for reviews with self-attention based framework using raw review texts. And our collected dataset contains the rating score of each review and the final decision of each paper.

2.2 Attention Mechanism

Attention mechanism was proposed by researchers to improve the performance of different NLP tasks. There are two common attention mechanisms: additive attention (Bahdanau et al., 2015) and multiplicative attention (Rush et al., 2015; Vaswani et al., 2017; Peng et al., 2019), they use different compatibility functions to compute the attention weights. Lin et al. (2017) introduced self-attention to extract an interpretable sentence embedding. Yang et al. (2016) proposed a hierarchical attention network for document classification, which applied attention mechanism at word and sentence level. Vaswani et al. (2017) built a simple network architecture based only on attention mechanism without convolutions and recurrence. Yin and Schütze (2018) proposed an attentive convolution network which enables deriving higher-level features for a word from information extracted from nonlocal context. Shen et al. (2018a) designed a new attention mechanism which is directional and multi-dimensional, and a neural network solely based on this attention mechanism was proposed to learn sentence embedding. Shen et al. (2018b) proposed a memory-efficient bi-directional self-attention network which splits sequence into blocks to save memory.

Our framework is also based on self-attention mechanism, which makes use of the hierarchical characteristic of HAN (Yang et al., 2016) and the ability of capturing relationships between words from two directions in DiSAN (Shen et al., 2018a).

3 Methodology

In this section, we first describe the problem setting, and then present the details of our proposed framework for paper review rating prediction and recommendation.

3.1 Problem Setting

We consider the problem of paper review rating prediction and recommendation from a dataset containing K papers, where each paper has M reviews associated with the corresponding ratings and a decision class. Concretely, given the set $R = \{(r_1, c_1), ..., (r_M, c_M), y\}$ for a scientific paper, where r_i is the i-th reviewer's text review and c_i is its associated numeric rating and y is the final decision (i.e., accept or reject). Assume that each text review r_i has N sentences $S = \{\mathbf{s}_{i,1}, \mathbf{s}_{i,2}, \cdots, \mathbf{s}_{i,N}\}$ and each sentence contains L words, let $\mathbf{w}_{i,j,t}$ with $i \in [1, M], j \in [1, N], t \in [1, L]$ denotes the t-th word in the j-th sentence of the i-th review document. Given a new paper with a set of reviews $R = \{r_1, ..., r_M\}$, our goal is to predict the decision class y which enables the program chairs to automatically make the final decision/recommendation, and also generate a rating c for each review r that is consistent with text sentiment as an aid to reviewers for discovering the rating inconsistency between ratings and review sentiments in the review process. Similar to (Zhang et al., 2010; Hassan and Shoaib, 2020), here we treat paper review rating prediction problem as a multi-class classification problem, where the class labels are the rating scores c. We treat the final decision prediction as a binary classification problem, where the class labels are the decisions y.

3.2 Our Approach

The proposed framework takes raw review texts as input and mainly consists of four components: sentence encoder, intra-review encoder, inter-review encoder and rating predictor, as shown in Figure 1. Before describing the details of each component, we introduce the multi-dimensional source2token self-attention module by following (Shen et al., 2018a) and taking this module in the sentence encoder as

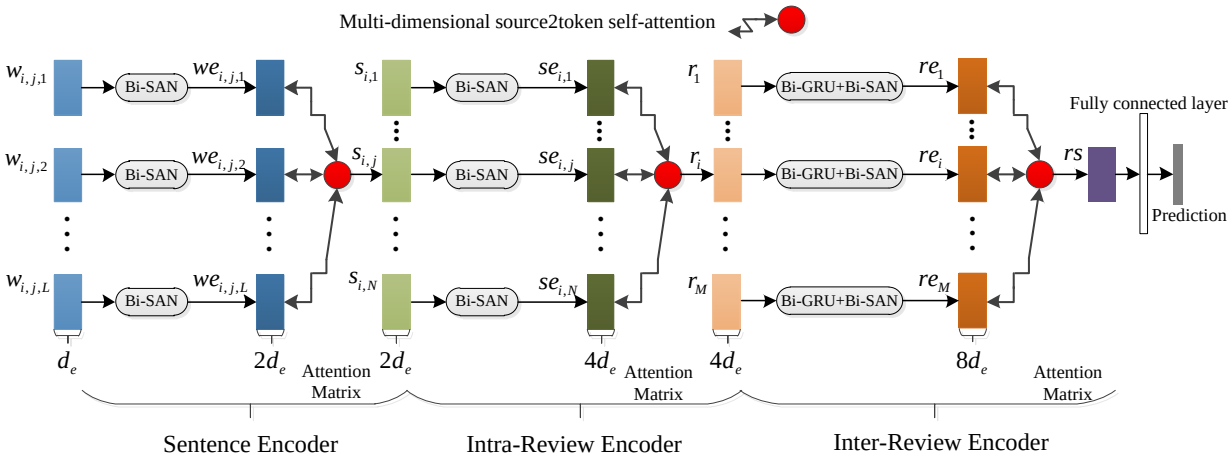


Figure 1: The architecture of HabNet framework.

an example. The attention weight of each word $\mathbf{we}_{i,j,t}, t \in [1, L]$ is obtained by applying softmax on the scores $f(\mathbf{we}_{i,j,t}), t \in [1, L]$ calculated by Eq. (1), $W^T, W^{(1)}, b^{(1)}, b$ are trainable parameters. The output of this module is the weighted sum of the inputs (e.g., $\mathbf{we}_{i,j,t}, t \in [1, L]$ in sentence encoder).

$$f(\mathbf{we}_{i,j,t}) = W^T \sigma(W^{(1)} \mathbf{we}_{i,j,t} + b^{(1)}) + b. \tag{1}$$

• ***Sentence Encoder.*** Sentence encoder is designed to capture the relationships between words in a sentence and the importance of each word to the meaning of that sentence. It is shown in the first part of Figure 1. It first generates context-aware embedding for each word in a sentence by using bi-directional self-attention module (Bi-SAN) (Shen et al., 2018a). Based on these context-aware embeddings of words, the encoding for that sentence, which contains all words' information and relations between words, is then obtained from the multi-dimensional source2token self-attention module (Shen et al., 2018a) which aims at generating the sentence encoding by combining the context-aware word embeddings. Specifically, the input of sentence encoder are pre-trained word embeddings obtained from raw review texts by using GloVe pre-trained word embedding (Pennington et al., 2014), or using BERT (Devlin et al., 2019) or SciBERT (Beltagy et al., 2019). Each word (e.g., $\mathbf{w}_{i,j,1}$, $\mathbf{w}_{i,j,2}$) is represented by a d_e-dimensional vector. These vectors are fed into Bi-SAN, which includes a forward self-attention network and a backward self-attention network. Each of these two networks outputs a refined embedding for each word and then the two refined embedding of each word are concatenated by Bi-SAN as the final context-aware embedding for each word (e.g., $\mathbf{we}_{i,j,1} \in \mathbb{R}^{2d_e}$). The context-aware embedding for each word has $2d_e$ dimension because of the two networks (i.e., forward and backward) in Bi-SAN. After obtaining the context-aware embedding of each word, sentence encoder can generate encoding $\mathbf{s}_{i,j} \in \mathbb{R}^{2d_e}$ for each sentence through the multi-dimensional source2token self-attention module.

• ***Intra-Review Encoder.*** Sentences in one review may have temporal orders, causality and other logic relationships, and some sentences contain more information for the review. Therefore, intra-review encoder is designed to capture these relations existing in each individual review itself. The input is the sentence embedding $\mathbf{s}_{i,j}$ generated by the first-level sentence encoder. The structure of intra-review encoder is similar to sentence encoder where it first feeds sentence embedding to the Bi-SAN module, which captures the relations between sentences and the importance of one sentence to another from two directions by generating forward embedding $\mathbf{s}_{i,j}^{fw}$ and backward embedding $\mathbf{s}_{i,j}^{bw}$ for sentence $\mathbf{s}_{i,j}$. The final embedding $\mathbf{se}_{i,j} \in \mathbb{R}^{4d_e}$ for each sentence $\mathbf{s}_{i,j}$ in i-th review is generated by concatenating $\mathbf{s}_{i,j}^{fw}$ and $\mathbf{s}_{i,j}^{bw}$. We have $\mathbf{se}_{i,j} = [\mathbf{s}_{i,j}^{fw} || \mathbf{s}_{i,j}^{bw}]$, where $||$ denotes concatenation operation. Next, the multi-dimensional source2token self-attention module takes $\mathbf{se}_{i,j}$ as input and generates encoding $\mathbf{r}_i \in \mathbb{R}^{4d_e}$ for i-th review by combining all $\mathbf{se}_{i,j}$ in this review according to their importance weights, i.e., attention weights. As shown in the second part of Figure 1, intra-review encoder can generate encoding $\mathbf{r}_i$ for each review of the same paper. The dimension of $\mathbf{r}_i$ is $4d_e$, which is double of sentence encoding because of Bi-SAN.

• ***Inter-Review Encoder.*** The integration of different reviews is essential for performing comprehensive analysis and supporting final decision-making on a paper. We use the inter-review encoder as the third level of our framework to integrate information from different reviews of each paper, as shown in the third part of Figure 1. It first feeds the second-level encoding $\mathbf{r}_i$ of i-th review of a paper to a bi-directional GRU (Bahdanau et al., 2015) layer, and then uses a Bi-SAN to model the relations between reviews from two directions by generating refined encoding $\mathbf{re}_i$ for this review. Thus $\mathbf{re}_i$ contains the information from other reviews. Then, a multi-dimensional source2token self-attention module is applied on these encoding $\mathbf{re}_i$ to get a final compact vector representation $\mathbf{rs}$ of the paper. This encoder can handle papers having different number of reviews by using padding. The whole process above is formulated as follows:

Step1: Feeding encoding $\mathbf{r}_i$ of each review of a paper generated by intra-review encoder to the bi-directional GRU layer, it outputs a new encoding for each review (we still use $\mathbf{r}_i$ to denote the new encoding of i-th review). Then these new encodings are fed to the following Bi-SAN module.

Step2: Bi-SAN has a forward self-attention network and a backward self-attention network. Two attention matrices, denoted as $\mathbf{P}^{i(fw)} \in \mathbb{R}^{4d_e \times M}$ and $\mathbf{P}^{i(bw)} \in \mathbb{R}^{4d_e \times M}$, for i-th review are calculated in these two networks respectively. Then the forward encoding $\mathbf{re}_i^{fw}$ and backward encoding $\mathbf{re}_i^{bw}$ for this review are generated as follows ($\odot$ denotes element-wise multiplication):

$$\mathbf{re}_i^{fw} = \sum_{o=1}^{M} \mathbf{P}_{\cdot o}^{i(fw)} \odot \mathbf{r}_o, \;\; \mathbf{re}_i^{bw} = \sum_{o=1}^{M} \mathbf{P}_{\cdot o}^{i(bw)} \odot \mathbf{r}_o, \tag{2}$$

where M is the number of reviews for one paper. $\mathbf{P}_{\cdot o}^{i(fw)}$ and $\mathbf{P}_{\cdot o}^{i(bw)}$ denote the o-th column in attention matrix $\mathbf{P}^{i(fw)}$ and $\mathbf{P}^{i(bw)}$ respectively. The refined encoding $\mathbf{re}_i$ for i-th review, which contains the information from other reviews of the same paper, is generated in the following equation.

$$\mathbf{re}_i = [\mathbf{re}_i^{fw} || \mathbf{re}_i^{bw}], \mathbf{re}_i \in \mathbb{R}^{8d_e}. \tag{3}$$

Step3: The multi-dimensional source2token self-attention module takes the encodings of all reviews of one paper outputed from Bi-SAN as input, and computes the importance weight for each review encoding $\mathbf{re}_i$, and then combines all these review encodings to get the final vector representation $\mathbf{rs}$ of the paper based on the importance weights in the similar way as shown in Eq. (2).

• ***Rating Prediction and Recommendation.*** With the three levels of encoding above, a fully connected layer with softmax function is designed to make rating prediction and final recommendation. Specifically, we take the compact representation $\mathbf{rs}$ from all reviews as its input to predict the final decision, and the encoding $\mathbf{r}_i$ for i-th review as its input to predict the corresponding rating, respectively. It is worth noting that the predicted review ratings are consistent with text sentiment conveyed by reviewers, thus it can serve as a guidance to reviewers for finding the inconsistencies between semantic review content and numerical review ratings in the review process.

3.3 Model Variants

To understand the contribution of different components in the proposed framework, we derive different variants for ablation study. Below are three variants implemented in our experiments.

HabNet-V1: After obtaining the encoding $\mathbf{r}_i$ of each review for a paper which is outputed from intra-review encoder, we sum them up using equal weight and then use the result as the final encoding $\mathbf{rs}$ of that paper, i.e., $\mathbf{rs} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{r}_i$. Thus the inter-review encoder is removed in this variant.

HabNet-V2: We remove the sentence encoder in the proposed framework as the second variant to verify the contribution of sentence encoder to the framework. Specifically, for a sentence, we use the average of all words' pre-trained embeddings as its encoding, and feed such sentence encodings to intra-review encoder. Therefore, this variant cannot encode the relations between words in a sentence.

HabNet-V3: We remove the intra-review encoder as the third variant to understand how effectively this encoder captures the interactions between sentences in a review document and to demonstrate the importance of intra-review encoder to the proposed framework. To be specific, the encoding of a review document is the mean of sentence embeddings in that review.

4 Experiments and Results

4.1 Dataset

We conduct experiments on two datasets to validate our approach for scientific paper decision recommendation and review rating prediction. One is called OpenReview dataset which is collected by us. The other one is a dataset extended from PeerRead which is originally published by Kang et al. (2018). Table 1 shows the statistics of OpenReview and Extended PeerRead. For all the experiments on these two datasets, all samples are randomly shuffled before splitting the dataset.

• ***OpenReview.*** This collection contains all reviews for ICLR conference and workshop's papers from 2017 to 2019. Generally, each paper has 3-5 reviews with corresponding ratings, the rating is a numeric value from 1 to 10 (10 being the highest rating). There is also a decision (accept or reject) associated with each paper. The number of accepted and rejected papers are 1341 and 1962. For paper decision recommendation, we use 2293, 491 and 492 papers as training, validation and testing set respectively. For review rating prediction, 7600, 1000, 1000 reviews are used as training, validation and testing set respectively. As shown in Table 1, the number of reviews with different ratings are highly imbalanced.

• ***Extended PeerRead.*** The majority of papers with reviews in the original PeerRead dataset are accepted papers collected from NIPS 2013-2017, as shown in Table 1, the number of accepted and rejected papers are 2054 and 0, respectively. Thus the original PeerRead dataset cannot be used directly for predicting final decisions on accepted/unaccepted papers due to the severe imbalance problem stated above. Therefore, we further collect 2211 papers from ICLR 2020 conference and corresponding reviews from the openreview website to extend the PeerRead dataset. Finally, the extended PeerRead dataset has 4265 papers and 13721 reviews in total. However, since most review ratings are not available in the original PeerRead, we only use this extended dataset to predict the final decision.

Dataset	Section	#Papers	#Reviews	#Acc/Rej	Review Rating Distribution									
					1	2	3	4	5	6	7	8	9	10
OpenReview	ICLR 2017-2019	3,303	9,600	1,341/1,962	37	205	851	1,816	1,943	2,154	1,875	563	158	16
Extended PeerRead	NIPS 2013-2017	2,054	7,006	2,054/0	-									
	ICLR 2020	2,211	6,715	687/1,524	-									
	Total	4,265	13,721	2,741/1,524	-									

Table 1: Statistics of OpenReview and Extended PeerRead, - means the rating distribution is unavailable.

4.2 Evaluation Metrics and Baselines

We use Accuracy, Macro-F1 and Micro-F1 to evaluate the effectiveness of our framework on the task of paper decision recommendation. For review rating prediction, due to the imbalanced distribution of ratings (shown in Table 1) and ineffectiveness of methods dealing with imbalanced problem (such as the oversampling technique and reducing rating range we tried), two new metrics with better discernibility are designed to better evaluate the performance of our framework and baselines apart from Accuracy.

Distance Measure (DM). The distance between true label and predicted label is crucial for evaluating a model when there are multiple labels as in our task of review rating prediction. The smaller the distance, the better a model works. Thus we design a new metric which incorporates the distance between predicted rating and true rating. This metric can distinguish a better model from a more reasonable perspective. For example, models which predict a rating of 8 as 7 are much better than models that predict it as 3. Let p_i and r_i be the predicted rating and true rating for i-th sample respectively, and n be the total number of samples. We define DM as follows:

$$DM = \frac{1}{n}\sum_{i=1}^{n}(1 - \frac{d_i}{d_{max}}), \quad \text{where} \quad d_i = |p_i - r_i|. \tag{4}$$

It first calculates the distance d_i for each sample and then takes an average over all samples according to Eq. (4). When the predictions for all samples are correct, the value of DM achieves its best which is 1. When all predictions are wrong and the distances between predicted ratings and true ratings are all

maximum distances d_{max} (in our case $d_{max} = 9$), the value of DM is 0. When the distances become smaller, the value of DM becomes larger. Thus it can evaluate the performance of models appropriately. The range of DM's value is $[0, 1]$. The larger its value is, the better the algorithm works.

Optimized Precision (OP). It is important to correctly predict all classes when the data is imbalanced. Inspired by Hossin and Sulaiman (2015), we combine accuracy and recall of all classes into a unified measure, which allows to better deal with imbalanced data environments. Let ACC be the accuracy, N be the number of classes, and R_i be the recall for i-th class, $i = 1, \cdots, N$. We define OP as follows:

$$OP = ACC - \frac{\sum_{i,j=1}^{N} |R_i - R_j|}{2(N-1)\sum_{k=1}^{N} R_k}. \tag{5}$$

As shown in Eq. (5), OP first computes the absolute differences between recalls of each pair of classes and sum them up, and then normalizes it by using the sum of all recalls. In this way, this metric measures the model's ability to predict the highest score of both accuracy and recall for all classes. The higher the value of OP, the better the model fits the data.

We compare our proposed method with fourteen other state-of-the-art methods and three variations of our proposed model: (1). 10 flat baselines: three are traditional text classification models, including Support Vector Machines (SVM), Logistic Regression (LR) and Naïve Bayes (NB); five are deep learning models, including RNN (i.e., Bi-GRU) (Cho et al., 2014; Bahdanau et al., 2015), TextCNN (Kim, 2014), TextRCNN (Lai et al., 2015), VDCNN (Conneau et al., 2017) and DPCNN (Johnson and Zhang, 2017); two are attention-based models, including Transformer (Vaswani et al., 2017) and SA-Sent-EM (Lin et al., 2017), which exploit various relationships existing in review text. (2). 4 hierarchical baselines which leverage the hierarchical structure of the dataset: HAN-extended is an extension of HAN (Yang et al., 2016) re-implemented by us; we also implement three Bert-based baselines (Devlin et al., 2019; Beltagy et al., 2019) using large pre-trained contextual embeddings. Specifically, Bert-base and Bert-large use 768 and 1024-dimensional Bert embedding respectively, while SciBert utilizes 768-dimensional SciBert embedding. For our proposed framework HabNet, apart from using GloVe embedding, we also conduct experiments by using the above three Bert-based contextual embeddings. (3). To demonstrate the contribution of each encoder ingredient, we also implement three variants of our proposed framework.

4.3 Experimental Settings

We use raw review texts as input for all models. For the decision recommendation task, 50-dimensional pre-trained GloVe word embedding is used for the models of our framework and HAN-extended, while 100-dimensional one is adopted for other deep learning models except bert-based baselines which utilize corresponding bert-embeddings. The number of training epochs is set to 100. For the rating prediction task, except bert-based baselines using bert embeddings, 100-dimensional pre-trained GloVe word embedding is used for all models. The number of training epochs is set to 50 since all models converge quickly. We use cross entropy as objective function to train all deep learning models. The common parameters, such as learning rate and batch size, are empirically set. For all the experiments on HabNet and its variants, we train each of them 10 times and use the average results to evaluate them.

4.4 Experimental Results and Discussion

The experimental results are shown in Table 2. For the paper decision recommendation task, HabNet achieves the best performance results no matter which kind of embedding is used. This demonstrates the effectiveness of our framework and its generality. To be specific, compared with flat baselines, our framework with GloVe embedding, i.e., HabNet (Glove), performs much better, which demonstrates that our framework can make good use of the hierarchical structure in the dataset. While compared with hierarchical baselines, HabNet (Bert-base), HabNet (Bert-large) and HabNet (SciBert) obtain good performance gain (5.4%, 8.2%, 5.4% and 8.2%, 8.0%, 4.1% in terms of accuracy on both datasets) over corresponding bert-based baseline respectively. The improvement indicates that our framework can capture the relationships between words, sentences, and reviews existing in the dataset. Although the three bert-based baselines can obtain contextual word embeddings, they cannot capture intra-review level

	Models	Scientific Paper Decision Recommendation						Rating Prediction		
		OpenReview			Extended PeerRead			OpenReview		
		ACC	Ma-F1	Mi-F1	ACC	Ma-F1	Mi-F1	ACC	DM	OP
Flat Baselines	NB	0.599	0.375	0.449	0.643	0.391	0.504	0.225	0.854	-0.772
	LR	0.699	0.635	0.671	0.794	0.763	0.789	0.316	0.883	-0.401
	SVM	0.686	0.603	0.643	0.790	0.757	0.783	0.318	0.883	-0.392
	RNN (Bahdanau et al., 2015)	0.606	0.404	0.472	0.697	0.628	0.679	0.250	0.856	-0.318
	TextCNN (Kim, 2014)	0.677	0.606	0.638	0.820	0.802	0.821	0.284	0.861	-0.356
	TextRCNN (Lai et al., 2015)	0.648	0.595	0.624	0.816	0.803	0.819	0.271	0.854	-0.364
	VDCNN (Conneau et al., 2017)	0.616	0.551	0.584	0.667	0.456	0.565	0.233	0.849	-0.350
	DPCNN (Johnson and Zhang, 2017)	0.642	0.478	0.551	0.831	0.828	0.835	0.295	0.874	-0.230
	Transformer (Vaswani et al., 2017)	0.602	0.381	0.458	0.720	0.658	0.705	0.212	0.841	-0.322
	SA-Sent-EM (Lin et al., 2017)	0.699	0.662	0.683	0.831	0.821	0.834	0.323	0.885	-0.204
Hierarchical Baselines	HAN (Yang et al., 2016)-extended	0.713	0.709	0.716	0.833	0.816	0.834	0.338	0.887	-0.187
	Bert-base (Devlin et al., 2019)	0.735	0.702	0.721	0.814	0.806	0.817	0.331	0.887	-0.208
	Bert-large (Devlin et al., 2019)	0.736	0.706	0.725	0.816	0.803	0.816	0.330	0.884	-0.270
	SciBert (Beltagy et al., 2019)	0.746	0.730	0.743	0.845	0.831	0.844	0.341	0.887	-0.176
Ours	HabNet (Glove)	**0.753**	**0.730**	**0.745**	**0.876**	**0.863**	**0.877**	**0.356**	**0.890**	**-0.061**
	HabNet (Bert-base)	**0.775**	**0.766**	**0.776**	**0.881**	**0.870**	**0.880**	**0.375**	**0.901**	**0.019**
	HabNet (SciBert)	**0.786**	**0.779**	**0.787**	**0.880**	**0.869**	**0.879**	**0.365**	**0.899**	**-0.013**
	HabNet (Bert-large)	**0.796**	**0.787**	**0.796**	**0.881**	**0.873**	**0.882**	**0.379**	**0.900**	**0.032**

Table 2: Performance results of all models on OpenReview and Extended PeerRead datasets. "ACC", "Ma-F1" and "Mi-F1" denote Accuracy, Macro-F1 and Micro-F1 respectively.

and inter-review level relationships as our framework does. Even HabNet (Glove) still performs better than the three bert-based baselines using pre-trained contextual embedding and HAN-extended, which further demonstrates the ability of the encoders on capturing the three-level relationships. In addition, the best performance of our framework on both datasets validates its generality, and HabNet with different embedding outperforming all the baselines consolidates this. It is worth noting that the performance results of all models on the extended PeerRead dataset are higher than that on the OpenReview dataset. The reason may be that the review texts (especially those from NIPS 2013-2017) in the PeerRead dataset are much shorter and less complex than those in the OpenReview dataset. This demonstrates that our framework can work on long review text much better than other models.

Note that in Table 2, there are only results on OpenReview dataset for review rating prediction, as PeerRead does not contain ratings. HabNet with various embedding (including GloVe and bert embeddings) achieving the best performance demonstrates the effectiveness and generalization ability of our framework again, because HabNet has a similar performance improvement as in the paper decision recommendation task when compared with flat and hierarchical baselines. Furthermore, the ratings predicted by HabNet, although not completely correct, can still be used as an aid to find inconsistencies between given ratings and text sentiments conveyed by reviewers.

4.5 Ablation Study

We conduct ablation study of our framework to evaluate the contribution of each component. The results are shown in Table 3. For the paper decision recommendation task, the better performance of HabNet over HabNet-V1 on both datasets indicates that the inter-review encoder can integrate information from different reviews of one paper well which helps the decision recommendation. While HabNet performs better than HabNet-V2 verifies the importance of sentence encoder which can encode the relationships between words in a sentence. And HabNet outperforming HabNet-V3 demonstrates the ability of intra-review encoder to capture sentence-level relations in a review text and that such relations between sentences contribute much information to the meaning of the review document. The results of HabNet and the variants on the review rating prediction task have a similar trend which further validates the contribution of different encoders to the framework. In conclusion, the three encoders help HabNet capture three-level relationships in the dataset which plays a vital role on improving prediction performance.

Task	Models	OpenReview Dataset					Extended PeerRead Dataset		
		Accuracy	Macro-F1	Micro-F1	DM	OP	Accuracy	Macro-F1	Micro-F1
Decision Prediction	HabNet-V1	0.735	0.705	0.723	-	-	0.858	0.846	0.860
	HabNet-V2	0.736	0.716	0.730	-	-	0.861	0.843	0.860
	HabNet-V3	0.726	0.700	0.717	-	-	0.859	0.846	0.861
	HabNet	**0.753** ↑	**0.730** ↑	**0.745** ↑	-	-	**0.876** ↑	**0.863** ↑	**0.877** ↑
Rating Prediction	HabNet-V2	0.335	-	-	0.886	-0.218	-	-	-
	HabNet-V3	0.336	-	-	0.887	-0.210	-	-	-
	HabNet	**0.356** ↑	-	-	**0.890** ↑	**-0.061** ↑	-	-	-

Table 3: Results of ablation study of our framework on OpenReview and Extended PeerRead datasets. Accuracy, Macro-F1 and Micro-F1 are the metrics used for the decision prediction/recommendation task; while Accuracy, DM and OP are used for the rating prediction task, and there are no results for this task on the extended PeerRead dataset because this dataset does not have rating for each review. HabNet achieves the best results which are in bold, arrow ↑ indicates statistical significance ($p < 0.05$).

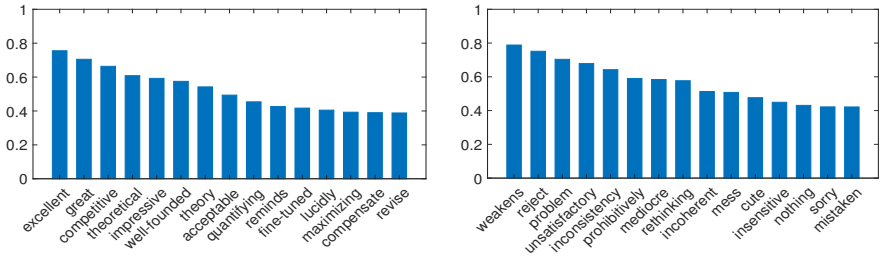


Figure 2: Attention weights of the top 15 approbatory words for accepted and rejected papers, left subfigure is for accepted papers, right one is for rejected papers.

4.6 Case Study

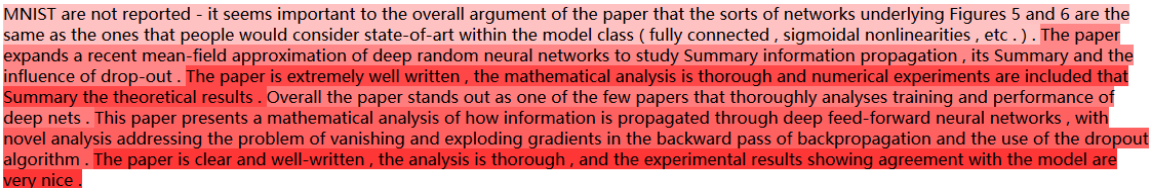
To gain a view over the ability of our proposed framework on capturing the importance of words in the scientific paper decision recommendation task, we visualize the top-15 approbatory words for accepted and rejected papers, as shown in Figure 2. One can see that for the accepted papers, the attention on positive words such as "excellent", "competitive" and so on are much higher than other words. For the rejected papers, negatives words such as "unsatisfactory", "incoherent" and so on have higher attention weight than others. Intuitively, reviewers can express their tendency towards the result of article more clearly through the above keywords. Moreover, compared to the attention weights of words in accepted papers, the attention weights of words in rejected papers are generally greater. The possible reason is that reviewers' comments on the rejected articles are more consistent than that of accepted articles.

We also visualize the sentence-level attention on one accepted paper and rejected paper respectively, as shown in Figure 3, the deeper the color, the bigger the attention weight. For the accepted paper, the sentence with the deepest color expresses strong positive attitude towards the acceptance decision, while other sentences without strong sentiment have smaller attention weights (i.e., the color is much lighter). The same trend also appears in the rejected paper . This result shows that our framework can capture the most important sentence-level signal within a review for predicting the final decisions for papers.

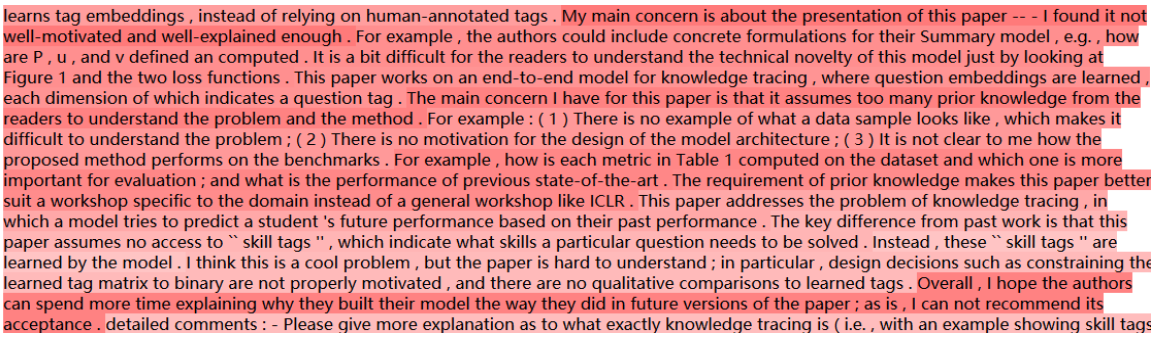
4.7 Error Analysis

We investigate the error cases that HabNet did not predict correctly on OpenReview dataset and find that: (1) 67% of them are predicted by HabNet as rejected papers, but they are actually accepted paper; (2) 33% of them are predicted as accepted but are in fact rejected.

We randomly select 20 examples from (1) and read the review text carefully, we find that: a). 18 of them have many negative keywords and phrases, like "unclear", "limited", "hard to interpret", "not provable" and so on. Although there are also positive expressions such as "looks good", "interesting", majority of the content contains negative ones. Thus the overall sentiment of the review text is classified as negative by HabNet. b). 2 of them have indicator of acceptance, like the keyword "accepted" or



(a) A review for an accepted paper.



(b) A review for a rejected paper.

Figure 3: Sentence-level attention visualization for accepted and rejected papers.

"acceptance", but meanwhile there are also many negative words, such as "confusing", "no comparison". All of these positive and negative information together make the model unable to make correct prediction.

We also randomly select 20 examples from (2). There are three cases: a). 7 of them contain very strong acceptance keywords and sentences, such as "pretty impressive", "promising", "I recommend acceptance". Because of these strong indicators, HabNet predict them as accepted papers. b). 2 of them have strong indicators of rejection, such as "The novelty of the paper is not enough to justify its acceptance", but they also have several strong positive keywords which deviate the overall sentiment of the review text and thus affect HabNet's prediction. c). 11 of them have many positive and negative keywords and sentences at the same time, and there is no strong indicator of rejection. HabNet can not deal with them very well, because it takes all the positive and negative information into consideration.

5 Conclusion

In this paper, a scientific paper review dataset called OpenReview is collected from ICLR openreview website and released. We observe that there is a three-level hierarchical structure in this dataset (i.e., word level, intra-review level and inter-review level) – the information and relationships between reviews of one paper may affect the final decision, and so may relationships between words and sentences in each review. Based on these observations, a hierarchical bi-directional self-attention network (HabNet) framework is proposed for paper review rating prediction and recommendation that can model the interactions among words, sentences, intra- and inter-reviews in an end-to-end manner. Moreover, considering the imbalanced distribution of different classes (i.e., ratings from 1 to 10) in the review rating prediction task, we design two new metrics to better evaluate models. It is seen that both experimental results of predicting final decisions for submitted papers and identifying ratings for reviews on two datasets (OpenReview and extended PeerRead) demonstrate our proposed framework has sufficient ability to capture the hierarchical structures of words, sentences and reviews in the datasets and outperforms other models. In the future, we plan to investigate multi-task learning for paper review rating recommendation.

Acknowledgements

The corresponding author is Hao Peng. This work is supported by the NSFC NO.62002007 and 61872022, NSF of Guangdong Province (2017A030313339), and in part by NSF under grants III-1763325, III-1909323, and SaTC-1930941. We thank the reviewers for their constructive comments.

Appendix A. Additional Details of Experimental Setting

All the experiments are conducted on GPU devices. The software platforms are Python 3.6.8 and Tensorflow 1.13.1.

Appendix B. Additional Experimental Results

For the review rating prediction task, we provide additional experimental results on OpenReview dataset under Macro-F1 and Micro-F1 metrics. The results are shown in Table 4. Our proposed framework still achieves the best results (in bold) on Macro-F1 and Micro-F1, it outperforms all the baselines in a similar trend as under other metrics such as Accuracy, DM and OP.

	Models	Review Rating Prediction	
		OpenReview	
		Macro-F1	Micro-F1
Flat Baselines	NB	0.037	0.083
	LR	0.138	0.280
	SVM	0.142	0.283
	RNN (Bahdanau et al., 2015)	0.109	0.219
	TextCNN (Kim, 2014)	0.121	0.245
	TextRCNN (Lai et al., 2015)	0.134	0.250
	VDCNN (Conneau et al., 2017)	0.102	0.193
	DPCNN (Johnson and Zhang, 2017)	0.137	0.262
	Transformer (Vaswani et al., 2017)	0.040	0.085
	SA-Sent-EM (Lin et al., 2017)	0.182	0.310
Hierarchical Baselines	HAN (Yang et al., 2016)-extended	0.167	0.311
	Bert-base (Devlin et al., 2019)	0.188	0.297
	Bert-large (Devlin et al., 2019)	0.182	0.300
	SciBert (Beltagy et al., 2019)	0.195	0.311
Ours	HabNet (Glove)	**0.205**	**0.338**
	HabNet (Bert-base)	**0.223**	**0.330**
	HabNet (SciBert)	**0.216**	**0.325**
	HabNet (Bert-large)	**0.226**	**0.339**

Table 4: Additional results of review rating prediction on OpenReview dataset under Macro-F1 and Micro-F1 metrics.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the ICLR*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the EMNLP*, pages 3606–3611.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *TPAMI*, 35(8):1798–1828.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the EMNLP*, pages 1724–1734.

Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2017. Very deep convolutional networks for text classification. In *Proceedings of the EACL*, pages 1107–1116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL*, pages 4171–4186.

Wenliang Gao, Naoki Yoshinaga, Nobuhiro Kaji, and Masaru Kitsuregawa. 2013. Modeling user leniency and product popularity for sentiment classification. In *Proceedings of the IJCNLP*, pages 1107–1111.

Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. 2019. Does my rebuttal matter? insights from a major nlp conference. In *Proceedings of the NAACL*, pages 1274–1290.

Junaid Hassan and Umar Shoaib. 2020. Multi-class review rating classification using deep recurrent neural network. *NPL*, 51(1):1031–1048.

Mohammad Hossin and MN Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. *IJDKP*, 5(2):1.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. In *Proceedings of the NAACL*, pages 2131–2137.

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the ACL*, pages 562–570.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. In *Proceedings of the NAACL*, pages 1647–1661.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the EMNLP*, pages 1746–1751.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Proceedings of the AAAI*, pages 2267–2273.

Youfang Leng, Li Yu, and Jie Xiong. 2019. Deepreviewer: Collaborative grammar and innovation neural network for automatic paper review. In *Proceedings of the ICMI*, pages 395–403.

Siqing Li, Wayne Xin Zhao, Eddy Jing Yin, and Ji-Rong Wen. 2019. A neural citation count prediction model based on peer review text. In *Proceedings of the EMNLP*, pages 4916–4926.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *Proceedings of the ICLR*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, pages 115–124.

Hao Peng, Jianxin Li, Yu He, Yaopeng Liu, Mengjiao Bao, Lihong Wang, Yangqiu Song, and Qiang Yang. 2018. Large-scale hierarchical text classification with recursively regularized deep graph-cnn. In *Proceedings of the WWW*, pages 1063–1072. International World Wide Web Conferences Steering Committee.

Hao Peng, Jianxin Li, Senzhang Wang, Lihong Wang, Qiran Gong, Renyu Yang, Bo Li, Philip Yu, and Lifang He. 2019. Hierarchical taxonomy-aware and attentional graph capsule rcnns for large-scale multi-label text classification. *TKDE*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the EMNLP*, pages 1532–1543.

Chao Qiao, Bo Huang, Guocheng Niu, Daren Li, Daxiang Dong, Wei He, Dianhai Yu, and Hua Wu. 2018. A new method of region embedding for text classification. In *Proceedings of the ICLR*.

Lizhen Qu, Georgiana Ifrim, and Gerhard Weikum. 2010. The bag-of-opinions method for review rating prediction from sparse text patterns. In *Proceedings of the COLING*, pages 913–921. Association for Computational Linguistics.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the EMNLP*, pages 379–389.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. 2018a. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *Proceedings of the AAAI*, pages 5446–5455.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang. 2018b. Bi-directional block self-attention for fast and memory-efficient sequence modeling. In *Proceedings of the ICLR*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the NIPS*, pages 5998–6008.

Baoxin Wang. 2018. Disconnected recurrent neural networks for text categorization. In *Proceedings of the ACL*, pages 2311–2320.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the NAACL*, pages 1480–1489.

Pengcheng Yang, Xu Sun, Wei Li, and Shuming Ma. 2018. Automatic academic paper rating based on modularized hierarchical convolutional neural network. In *Proceedings of the ACL*, pages 496–502.

Wenpeng Yin and Hinrich Schütze. 2018. Attentive convolution: Equipping cnns with rnn-style attention mechanisms. *TACL*, 6:687–702.

DongMei Zhang, Shengen Li, Cuiling Zhu, Xiaofei Niu, and Ling Song. 2010. A comparison study of multi-class sentiment classification for chinese reviews. In *Proceedings of the FSK*, volume 5, pages 2433–2436. IEEE.