





An Effective Chinese Text Classification Method with Contextualized Weak Supervision for Review Autograding

Yupei Zhang^{1,2}(✉) , Md Shahedul Islam Khan¹ , Yaya Zhou^{1,2}, Min Xiao³,
and Xuequn Shang^{1,2}(✉)

¹ School of Computer Science, Northwestern Polytechnical University, Xi'an, China
{ypzhaang, shang}@nwpu.edu.cn

² MIT Big Data Storage and Management Libraries, Xi'an, China

³ Graduate School, Northwestern Polytechnical University, Xi'an, China

Abstract. This paper aims to develop a Chinese text classification workflow in education situations, where a grade can swing due to subjective cognitive loads. This problem is often observed between the academic paper comments and their grades, leading to a challenge in Chinese texts. To analyze this problem, we in this paper introduce an effective Chinese text classifier by extending the popular seed words-based model into an effective workflow. We first made texts into vectors in the proposed method using Chinese preprocessing. We then exploited the bidirectional encoder representations from Transformers to integrate the contextualization features, then performed a hierarchical attention network for classification. In this study, we collected 4,310 review comment short-texts involving 140 universities in China. As these texts include noisy grades from experts, the proposed method yields seed words for each category, resulting in pseudo labels to weakly supervise the network training instead of the noisy labels. We finally evaluated the designed workflow on the real-world datasets and achieved a good performance in Chinese classification compared with the traditional models. This study provides insights into a real educational text case where a review grade can swing due to subjective cognitive loads and an available workflow to automatically grade these Chinese expert comment texts, facilitating the precise academic evaluation system.

Keywords: Educational data mining · Text data analysis · Weakly supervision · BERT · Chinese text classification · Review auto grading

1 Introduction

In recent years, introducing the weak supervision for text classification has been vastly popular in researches because labels are often lacking in practice, resulting in many rigid techniques and tools for text classification with weak tags [1, 2]. Researchers have already addressed many educational problems by using the powerful machine learning models, such as knowledge diagnosis [3, 5], learning performance prediction [4, 6],

path optimization recommendation [7]. However, associations between Chinese academic review comments and the subsequent grades have not been touched upon. These reviews are generally used to evaluate if a graduate student can gain a master's degree or not and control the quality of graduate students. This problem falls into the short-text classification category, which is a fundamental task in natural language processing (NLP) [8] and has been studied in many applications including automated customer relationship management [9], paragraph summarization [10], and question answer system [11].

In general, commonly used text classification methods are comprised of two main tasks, i.e., feature engineering and classification. Feature engineering takes the primary steps to obtain word vectors, including data cleaning, unnecessary characters and words removal, and performing vectorization by leveraging TF-IDF (Term Frequency-Inverse Document Frequency), Word2Vec, or GloVe (Global Vector for Word Representation). Afterwards, classification techniques are implied for classification tasks, e.g., support vector machine [12] and XGboost [13].

The nature of human language often challenges the traditional methods of text representations by high sparsity. As a result, most existing models usually ends in simple features while incurring computation costs. On the other hand, the ambiguous nature of human language toughens the procedure of feature extraction from texts. On top of that, it is difficult to acquire sufficient labeled texts data. To resolve these issues, many state-of-the-art extremely performant Deep Learning NLP techniques, e.g., BERT [14], ERNIE [15] and GPT [16], have been proposed to gain near-supervised accuracy in text classification through exploiting the weakly-supervised techniques. Yosi et al. proposed an end-to-end document retrieval system using a weak supervised deep model with BERT and GPT2 [17]; Zihan et al. came up with a weakly-supervised text classification method based on Key Word graphs [18].

While weakly supervised text classification has been researched widely, it has been overlooked in Chinese educational texts. In this study, we propose a Chinese text classification workflow to learn the relation between expert-provided comments and their annotated grades for academic articles. More specifically, this study introduces a Chinese short-text classifier by extending the popular seed words-based model [23] into a systemic workflow. We first made texts into vectors in our proposed method using Chinese preprocessing and then exploited the bidirectional encoder representations from Transformers [19] to integrate the contextualization features, followed by a hierarchical attention network for performing text classification. Our method can address the existing conflict between expert given comments and their grades and provides better classification accuracy. Altogether, automatic grading of the expert given reviews would progress the educational evaluation process by manifolds.

The remaining of our paper is organized to (1) introduce the data and the proposed workflow in Sect. 2; (2) Provide and analyze the experiment results in Sect. 3; and (3) Conclude our study and discuss the future works in Sect. 4.

In this study, we provided a thorough comparison to portray the superiority of our method over traditionally used linear models like SVM, NB, and nonlinear methods like CNN.

2 Methodology

2.1 Dataset Collection

We have collected academic dissertation reviews from over 140 Chinese universities for this study. The data includes 4,310 reviews, comprising over six years (2014–2020) span.

Table 1. The used Chinese dissertation review dataset

Dataset	Number
Document	4296
Average document len	209
Grade category	5

These reviews contain basic information about the students, academic graduation thesis title, evaluation reviews and revision, expert provided grades, etc. Usually, the evaluation reviews and the revision comments lead to the potential expert-provided grades, divided into five categories, i.e., Excellent, Good, General, Bad, and Poor. We use the English translations of the assigned grades instead of Chinese, as shown in Fig. 1.

Table 2. The class distribution in our dataset

Grading class	Available data
General	1184
Excellent	440
Poor	128
Well	2292
Bad	252

With regards to our approach, we removed the non-graded reviews that led us to have a total of 4296 data for our study.

We had a fair amount of data for this study, but the data distribution among the classes were highly imbalanced, which can be observed in Fig. 1. Table 1 summarizes the dataset’s statistics. Table 2 summarizes the data available for each class in our final dataset.

We anticipated that this data was noisy due to the fact that the expert-assigned grade fluctuated according to the individual’s perspective. This assumption was also validated by the model’s performance. Our previous paper demonstrated this phenomenon [20].

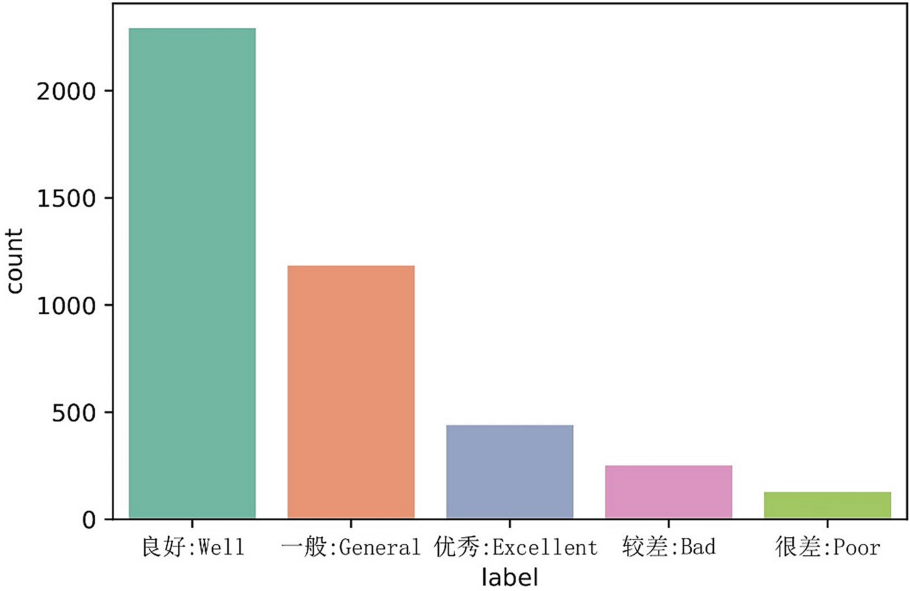


Fig. 1. Data distribution in the review category.

2.2 The Proposed Workflow

To design our workflow, we have developed a data preparation system that preprocesses Chinese text data and prepare the data-frame to be used in our selected model. Following the text preprocessing system, we have selected a seed word-based contextualized weak supervision model, “ConWea” [23] which leverages use of contextualized representation techniques to generate contextualized vectors for each word occurrence. However, in our experiments, we used “Bert-base-Chinese” [24] for Chinese text embeddings. Afterwards, the model generates pseudo labels and trains a neural text classifier on those labels along with the contextualized corpus while expands initial seed words list based on a ranking system it employs.

We feed our collected text data to our workflow along with the initial seed words. To begin, we pass the collected Chinese review texts through the preprocessing steps to clean it up and make the data-frame model [23] ready. Then the model ready data-frame and the seed words are provided to the model. The model disambiguates the seed words by explicitly learning different senses (meanings) of each word with contextualized word embeddings. It first performs k-means clustering for each word in the vocabulary to identify potentially different senses (meanings), then eliminates the ambiguous keyword senses leading to a fully contextualized corpus. This contextualized corpus and generated pseudo labels trains the hierarchical attention network classifier and employing a ranking system extends the initial given seed words and iteratively uses these seed words alongside the contextualized corpus.

Figure 4, illustrates the whole method and shows the data flow through our proposed workflow.

Data Preparation. To begin, we addressed the word segmentation issue of our collected text data as Chinese texts lacks proper word to word separation. We used “jieba” [21], a natural language toolkit that works well with the Chinese texts to obtain proper word segmentation.

Figure 2 shows the unsegmented data and Fig. 3 shows the properly segmented Chinese texts in our dataset.

sentence	Label
论文选题能够契合物联网和云计算发展的需求，有实践意义。论...	良好
论文选题有明确的国防应用背景，采用数值模拟方法进行研究亦...	一般
论文介绍了天津涉海企业经济数据监测系统的设计与实现，有一...	较差
该论文类似于工作报告和项目汇报，内容是所开发系统的主要功...	较差
论文采用数值模拟与试验相结合的方法对某飞机大偏距双通道S...	良好
现代飞机结构设计中，高温引起的稳定性问题成为是一个非常重...	一般

Fig. 2. Collected unsegmented text data

Our data cleansing process entails removing unnecessary characters and noise-introducing words, as well as adding certain words to the “jieba” dictionary., etc.

sentence	Label
陶瓷型芯制备是航空发动机气冷空心叶片精密铸...	良好
论文针对兖矿集团养老保险管理信息系统进行设计和...	较差
论文研究中信证券客户管理系统的项目进度管理，...	良好
论文选择建设生态文明建设背景下的地方政府的...	一般
该论文选题合理，为动密封结构的研究提供理论...	良好
论文针对高等学院校企合作管理系统进行设计和开...	优秀

Fig. 3. Word segmented Chinese review comments

We opted for “jieba” for tokenization and vocabulary generating. Our dataset generated a vocabulary length of 9927. For sentence segmentation task, we used “harvesttext” [22] which is another well-performing toolkit for natural language tasks on Chinese texts.

We used Chinese stop words list from a GitHub repository [31] and had it refined based on our corpus.

Chinese Corpus Contextualization. Following data preparation, the corpus and the seed words are provided to the selected model. We utilized a seed words-based contextualized weak supervision method [23], “ConWea” in the proposed workflow method, that leverages the use of contextualized representation techniques to generate contextualized vectors for each word occurrence. We fed the model with the corpus and initial

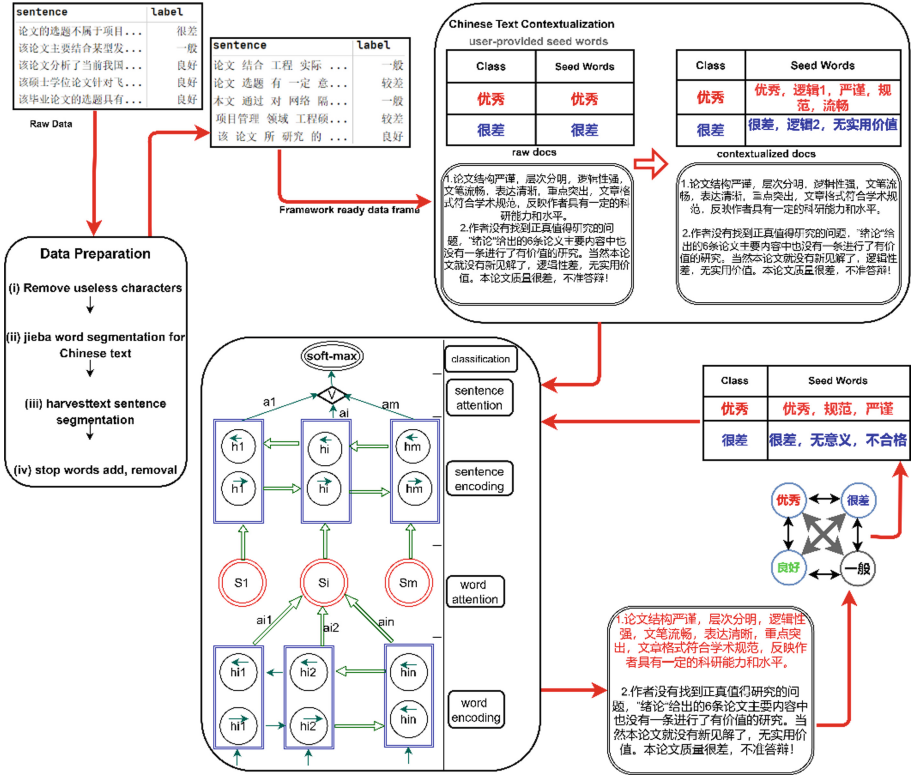


Fig. 4. The proposed workflow of Chinese auto-grading model training

seed words and obtained a contextualized corpus. In our experiments, we used “Bert-base-Chinese” to obtain text embeddings. Following data preparation, the corpus and the seed words are provided to the model. Afterwards, the model groups the word occurrences of the same word into an adaptive number of interpretations based on the label indicative seed-words, resulting in a contextualized corpus.

BERT for Chinese Text. BERT stands for bidirectional encoder representation from transformers which is a transformer-based machine learning technique that has been pre-trained by Google [14]. It was created to assist computers in understanding the meaning of ambiguous human language in text by utilizing surrounding texts and therefore comprehending context. BERT’s key technical feature is applying the bidirectional training of transformer, a popular attention model, to language modelling. This is in contrast to previous efforts which looked at a text sequence either from left to right or combined left-to-right and right-to-left training. Leveraging this capability, it is pre-trained on Masked Language Modeling [26] and Next Sentence Prediction. We use the version, “Bert-base-Chinese” [24], in our experiment as we are working with Chinese text to get contextualized vector representation of every word occurrence.

Classifier Training and Expanding Seed Words. Leveraging provided seed-words for each label, the model uses pseudo labels to train the contextualized Chinese corpus and train a neural classifier on those labels. We used the class names as initial seed-words.

In our experiments, we used a hierarchical attention network [25] that considers the hierarchical structure of the text data i.e., document-sentence-words, and integrates an attention mechanism that finds the most important words and sentences in a document considering the context. The architecture has been provided in Fig. 4.

There are two levels of attention whereas word level attention identifies the important word in sentence and sentence level attention does the same for the document. The contextualized corpus alongside the seed words is fed to the HAN model, and predicted pseudo-labels are used for document classification.

Leveraging contextualized Chinese corpus along with the predicted labels the model ranks the contextualized words and the top words are included in the seed words list per class. To get the top words, how label indicative the words are, frequency of that ideal seed words in the documents and unusualness of the words are considered. The seed words expansion and documents classification happen in an iterative manner inside “ConWea” [23] and the iteration number is controlled by T as a tunable hyper parameter in the model. We have set it to 7 as by then the method achieves convergence therefore the expanded seed sets and classification performance become constant.

2.3 Model Selection

On this dataset, we utilized three commonly used text classifiers. We implemented Linear Support Vector Machines (LSVM) and Multinomial Naive Bayes (MNB) as linear classifiers. Additionally, we used a nonlinear model based on the classical Convolutional Neural Network (CNN). We evaluated these models’ performance on our dataset and compared with the result our proposed workflow method. We used ReLU [27] as the activation function in our classical CNN design and ADAM [28] as the optimizer. For LSVM and MNB stochastic gradient descent (SGD) was used and we did appropriate parameters tuning.

2.4 Performance Evaluation Metrics

We implemented several validation metrics to observe the model performance in depth. We used accuracy score, macro-precision, macro-recall, macro-f1-score, weighted-precision, weighted-recall, and weighted-f1 score to properly examine the ability of the classifier. The accuracy rate is the aggregated accuracy score of a model.

Our classification task is a multi-class classification task; therefore, we calculate the True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN). As we have five classes to classify among, we can consider $x = a, b, c, d, e$ for the five classes. Therefore, precision (P_x) and Recall (R_x) per class are calculated by

$$P_x = \frac{TP_x}{(TP_x + FP_x)} \quad (1)$$

$$R_x = \frac{TP_x}{(TP_x + FN_x)} \quad (2)$$

For macro-precision (MR), macro-recall (MR), and macro- $(F1)$ -score we use the equations below

$$Mp = 15 \sum_{x=a}^e (P_x), MR = 15 \sum_{x=a}^e (R_x), MF1 = \frac{2 * Mp * MR}{(MP + MR)} \quad (3)$$

As our dataset is highly imbalanced, to get proper evaluation, we will compute weighted-precision (Wp), weighted-recall (WR) and weighted-f1-score ($WF1$) as

$$WP = \frac{\sum_{x=a}^e P_x * N_x}{\sum_{x=a}^e N_x}, WR = \frac{\sum_{x=a}^e R_x * N_x}{\sum_{x=a}^e N_x}, WF1 = \frac{2 * Wp * WR}{WP + WR} \quad (4)$$

where (N_x) indicates the number of total data samples in the x -th class.

3 Result Evaluation

3.1 Overall Results

Table 3 shows the classification accuracy of all the methods we have tried the dataset on. As can be seen, our proposed workflow yields about 20% better accuracy than the Linear Support Vector Machine classifier and 25% better accuracy than the Multinomial Naive Bayes classifier and 13% better performance on accuracy score than the nonlinear CNN classifier leveraging contextualization. Therefore, portraying superiority in classification task performance compared to the widely used traditional methods. Observations remain uniform across all the classification metrics.

Table 3. The prediction accuracy on our data set by using all mentioned methods

Classification metrics				
Method	MNB	LSVM	CNN	Our Method
ACC	0.68	0.73	0.80	0.93
Macro-Precision	0.79	0.84	0.69	0.90
Macro-Recall	0.46	0.52	0.77	0.89
Macro-F1	0.51	0.58	0.72	0.89
Weighted-Precision	0.73	0.75	0.81	0.93
Weighted-Recall	0.68	0.73	0.80	0.93
Weighted-F1	0.66	0.72	0.80	0.93

Figure 5 illustrates, the classification prediction accuracy of all the methods used in each category of our dataset. We can observe the uniform superiority of our proposed

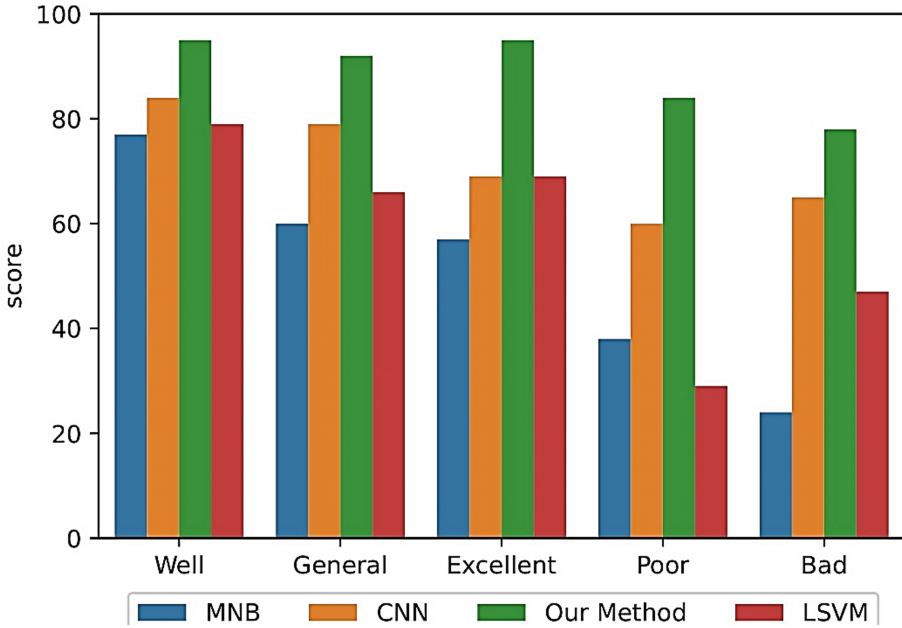


Fig. 5. The classification accuracy of MNB, CNN, Our Method, and LSVM

workflow method here over every other method. All of the approaches had difficulty distinguishing between the classes “Bad” and “Poor”. The reason behind this lies with the existing conflicts between experts given comments and their subsequent grades.

Table 4 explains the arbitrariness of assigning these two grades by the experts hence raising the conflict between the comments and the annotated grades that leads the workflow to fail to perform classification properly at times. Yet, the workflow could handle this issue better than other used methods. The substantial superiority of our approach is also demonstrated in classifying highly imbalanced datasets such as the one used, which has very little data in classes “Poor” (128) and “Bad” (252) compare to other classes.

Nonetheless, the classification accuracy of these classes is comparable to that of classes with more data. Other methods failed to accomplish this.

3.2 Confusion Matrix

Figure 6 shows the confusion matrix for the methods we have used in this experiment, LSVM, CNN, MNB, and our proposed workflow method for the classification task.

All of the approaches had difficulty distinguishing between the classes “Bad” and “Poor”, however our approach performed significantly better. The reason behind this lies with the existing conflicts between experts given comments and their subsequent grades. But our proposed workflow method shows significant superiority in addressing the issue.

Table 4. Example of arbitrariness in the expert annotated grades

Review Texts	Annotated Labels	Predicted Labels
论文以 Android	较差	很差
该学位论文描述一个采用 Power	很差	很差
论文针对 究矿 集团 养老保险 管理信息系统进行设计和开发，选题具有一定的实践意义	较差	较差
论文致力于研究视频监控系统，完成流媒体监控视频的归档和查询重放，有较好的技术和实用意义。论文首先分析了 RTSP	很差	较差

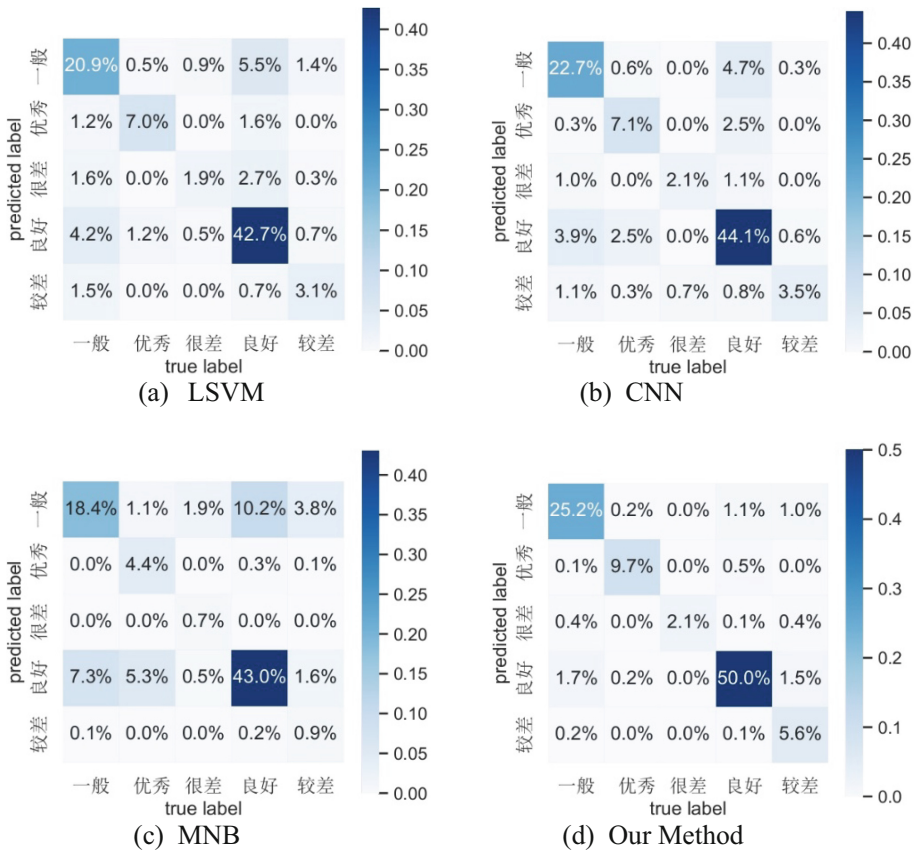


Fig. 6. Confusion Matrix of all the four methods used

4 Conclusion and Discussion

This study proposes a deep learning-based workflow method to explore the relation between expert-given comments and their subsequent grades in Chinese graduation thesis evaluation. The proposed workflow introduced an effective Chinese text classifier leveraging Chinese text preprocessing and then exploited one of the popular seed words-based weakly supervised methods [23] to create contextualized corpus followed by a hierarchical attention network.

We used 4296 evaluation reviews in our dataset that have been collected from over 140 higher educational institutions in China. Our proposed workflow method achieved.

93% accuracy, which is an improvement of 13% over classical CNN, 20% over LSVM and an astounding 25% over the NB-based method. In addition, our method performed uniformly well in classifying each grading class in terms of various implied metrics.

In the future, we will work on addressing the problem of noisy and imbalanced labels that exist at present in the educational situation and come up with scientific solutions for omitting those. More data analysis studies will also be developed, including visualization [29] and important factor discovery [30], linear dimensionality reduction of data [32], further works on unsupervised Order-Graph Regularized Sparse Dictionary Learning [33], students' knowledge diagnosis [34], prediction of undergraduate students learning performance [35], expanding research on deep learning based exploration in impact of tumor infiltrating immune cells [36], etc.

Acknowledgement. This study was funded in part by the National Natural Science Foundation of China (U1811262, 61802313, 61772426), the Key Research and Development Program of China (2020AAA0108500), the Reformation Research on Education and Teaching at Northwestern Polytechnical University (2021JGY31), the Higher Research Funding on International Talent cultivation at Northwestern Polytechnical University (GJGZZD202202), Research Topic at The Chinese Society of Academic Degrees and Graduate Education (2020ZA1008).

References

1. Wang, Y., Sohn, S., Liu, S., et al.: A clinical text classification paradigm using weak supervision and deep representation. *BMC Med. Inform. Decis. Mak.* **19**, 1 (2019)
2. Yu, M., Jiaming, S., Chao, Z., Jiawei, H.: Weakly-supervised neural text classification. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM 2018)*, pp. 983–992. Association for Computing Machinery, New York, NY, USA (2018)
3. Zhang, Y., Dai, H., Yun, Y., Liu, S., Lan, A., Shang, X.: Meta-knowledge dictionary learning on 1-bit response data for student knowledge diagnosis. *Knowl. Based Syst.* **205**, 106290 (2020)
4. Zhang, Y., An, R., Liu, S., Cui, J., Shang, X., 2021. Predicting and understanding student learning performance using multi-source sparse attention convolutional neural networks. *IEEE Trans. Big Data* 1–1 (2021)
5. Liu, Q., Shen, S., Huang, Z., Chen, E., Zheng, Y.: A survey of knowledge tracing. *arXiv preprint arXiv:2105.15106* (2021)

6. Yun, Y., Dai, H., Cao, R., Zhang, Y., Shang, X.: Self-paced graph memory network for student GPA prediction and abnormal student detection. In: Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., Dimitrova, V. (eds.) AIED 2021. LNCS (LNAI), vol. 12749, pp. 417–421. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-78270-2_74
7. Dwivedi, P., Kant, V., Bharadwaj, K.K.: Learning path recommendation based on modified variable length genetic algorithm. *Educ. Inform. Technol.* **23**(2), 819–836 (2017). <https://doi.org/10.1007/s10639-017-9637-7>
8. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J.: Deep learning-based text classification: a comprehensive review. *ACM Comput. Surv.* **54**(3), 1–40 (2021)
9. Mohamed, D.A.R., Sakre, M.M.: A performance comparison between classification techniques with CRM application. *SAI Intell. Syst. Conf.* **2015**, 112–119 (2015)
10. Kumar, G.K., Rani, D.M.: Paragraph summarization based on word frequency using NLP techniques. In: AIP Conference Proceedings, vol. 2317, p. 060001 (2021)
11. Anhar, R., Adji, T.B., Setiawan, N.A.: Question classification on question-answer system using bidirectional-LSTM. In: 2019 5th International Conference on Science and Technology (ICST), pp. 1–5 (2019)
12. En.wikipedia.org.: Support-vector machine – Wikipedia (2022). <https://en.wikipedia.org/wiki/Support-vector-machine>. Accessed 10 April 2022
13. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016), pp. 785–794. Association for Computing Machinery, New York, NY, USA (2016)
14. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR abs/1810.04805* (2018)
15. Yu, S., et al.: ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation (2021)
16. Shree, P.: The Journey of Open AI GPT models. *Medium* (2020). <https://medium.com/walmartglobaltech/the-journey-of-open-ai-gpt-models-32d95b7b7fb2>. Accessed 10 April 2022
17. Mass, Y., Roitman, H.: Ad-hoc document retrieval using weak-supervision with BERT and GPT2. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 4191–4197. Association for Computational Linguistics (2020)
18. Zhang, L., Ding, J., Xu, Y., Liu, Y., Zhou, S.: Weakly-supervised Text Classification Based on Keyword Graph (2021). <https://doi.org/10.18653/v1/2021.emnlp-main.222>
19. Wikimedia Foundation: Transformer (Machine Learning Model). Wikipedia (2022). Retrieved from 12 April 2022. [https://en.wikipedia.org/wiki/Transformer\(machine-learning-model\)32d95b7b7fb2](https://en.wikipedia.org/wiki/Transformer(machine-learning-model)32d95b7b7fb2). Accessed 10 April 2022
20. Zhang, Y., Zhou, Y., Xiao, M., et al.: Comment text grading for Chinese graduate academic dissertation using attention convolutional neural networks. In: 2021 7th International Conference on Systems and Informatics (ICSAI), pp. 1–6. IEEE (2021)
21. PyPI: jieba (2022). <https://pypi.org/project/jieba/>. Accessed 11 April 2022
22. Welcome to Harvesttext’s documentation: Welcome to HarvestText’s documentation - HarvestText 0.8.1.6 documentation. (n.d.). Retrieved from 11 April 2022. <https://harvesttext.readthedocs.io/en/latest/>. Accessed 11 April 2022
23. Mekala, D., Shang, J.: Contextualized weak supervision for text classification. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 323–333. Association for Computational Linguistics (2020)
24. Huggingface.co.: ckiplab/bert-base-chinese · Hugging Face (2022). <https://huggingface.co/ckiplab/bert-base-chinese>. Accessed 11 April 2022

25. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1480–1489. Association for Computational Linguistics, San Diego, California (2016)
26. Analytics India Magazine: A complete tutorial on masked language modelling using BERT (2022). <https://analyticsindiamag.com/a-complete-tutorial-on-masked-language-modelling-using-bert>. Accessed 14 April 2022
27. Agarap, A.F.: Deep learning using rectified linear units (relu). arXiv preprint [arXiv:1803.08375](https://arxiv.org/abs/1803.08375) (2018)
28. Diederik, K., Jimmy, B.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (2014)
29. Zhang, Y., Xiang, M., Yang, B.: Low-rank preserving embedding. *Pattern Recogn.* **70**, 112–125 (2017)
30. Zhang, Y., Xiang, M., Yang, B.: Hierarchical sparse coding from a Bayesian perspective. *Neurocomputing* **272**, 279–293 (2018)
31. Stopwords-Iso.: STOPWORDS-ZH/STOPWORDS-ZH.TXT at master · stopwords-ISO/stopwords-zh. GitHub (2020). Retrieved from 28 March 2022. <https://github.com/stopwords-iso/stopwords-zh/blob/master/stopwords-zh.txt>. Accessed 11 April 2022
32. Zhang, Y., Xiang, M., Yang, B.: Low-rank preserving embedding. *Pattern Recogn.* **70**, 112–125 (2017). ISSN 0031-3203. <https://doi.org/10.1016/j.patcog.2017.05.003>
33. Zhang Y, et al.: Multi-needle detection in 3D ultrasound images using unsupervised order-graph regularized sparse dictionary learning. *IEEE Trans. Med. Imaging* **39**(7), 2302–2315 (2020). <https://doi.org/10.1109/TMI.2020.2968770>. Epub 2020 Jan 22. PMID: 31985414; PMCID: PMC7370243
34. Zhang, Y., Dai, H., Yun, Y., Liu, S., Lan, S., Shang, X.: Meta-knowledge dictionary learning on 1-bit response data for student knowledge diagnosis. *Knowl. Based Syst.* **205**, 106290 (2020). ISSN 0950-7051. <https://doi.org/10.1016/j.knosys.2020.106290>
35. Zhang, Y., An, R., Liu, S., Cui, J., Shang, X.: Predicting and understanding student learning performance using multi-source sparse attention convolutional neural networks. *IEEE Trans. Big Data.* <https://doi.org/10.1109/TBDATA.2021.3125204>
36. Liu, S., Zhang, Y., Shang, X., Zhang, Z.: ProTICS reveals prognostic impact of tumor infiltrating immune cells in different molecular subtypes. *Brief Bioinform.* **22**(6), bbab164 (2021). <https://doi.org/10.1093/bib/bbab164>. PMID: 33963834